Environment Canada
Environnement Canada

Environment Canada
Environnement Canada
CANADA. INLAND WATERS DIRECTORATE.
SCIENTIFIC SERIES

Statistical Assessment of the Great Lakes Surveillance Program, 1966-1981, Lake Erie

Edited by A.H. El-Shaarawi

SCIENTIFIC SERIES NO. 136

NATIONAL WATER RESEARCH INSTITUTE
INLAND WATERS DIRECTORATE
CANADA CENTRE FOR INLAND WATERS
BURLINGTON, ONTARIO, 1984

85-1736/

Environment Canada
Environnement Canada

Statistical Assessment of the Great Lakes Surveillance Program, 1966-1981, Lake Erie

Edited by A.H. El-Shaarawi

SCIENTIFIC SERIES NO. 136

NATIONAL WATER RESEARCH INSTITUTE
INLAND WATERS DIRECTORATE
CANADA CENTRE FOR INLAND WATERS
BURLINGTON, ONTARIO, 1984

(Disponible en français sur demande)

© Minister of Supply and Services Canada 1984

Cat. No. En 36-502/136E

ISBN 0-662-13493-1

Contents

Page

ABSTRACT xix

RÉSUMÉ xxi

ACKNOWLEDGMENTS xxiii

1. INTRODUCTION 1

2. LAKE ERIE REVIEW 3
   Early historical work on dissolved gases 4
   Physical, biological and chemical variables controlling D.O. depletion. 6
   Present state of Lake Erie 8
   Conclusions 18

3. TEMPORAL CHANGES IN LAKE ERIE 27
   Introduction 27
   Data sources 29
   Statistical methods 30
      Reduction of spatial variability 30
      Determination of the seasonal cycle 32
      Detection of trend 34
   Water level 34
      A model for the water level 35
         The model 35
      Seasonal changes 39
   Niagara River flows 39
   Air temperature 40
   Water temperature 43
      The seasonal cycle of the surface water temperature 43
      Statistical analysis of the surface temperature data 44
      Year to year variation in the surface water temperature 46
      Vertical variation in temperature 47
      Year to year variability in the hypolimnion temperature (Central Basin) 48
   Chlorophyll a 49
      Western Basin 50
      Central Basin 52
      Eastern Basin 52
   Total phosphorus - TP 53
      Western Basin 53
      Central Basin 54
      Eastern Basin 54

Contents (Cont.)

Page

Soluble reactive phosphorus - SRP 54
Filtered nitrate nitrite (NO_3NO_2-N) 55
Ammonia (NH_3) 56
Secchi disk depth 57
Year to year variability in Secchi disk values . . . 58
Turbidity 58
Discussion of the results of fitting model 3.4 to the turbidity data 58
Year to year variability in turbidity values 59
Phosphorus and chlorophyll a association 59
Ratio of ammonia to nitrite and nitrate 60

4. SPATIAL AND TEMPORAL VARIABILITY OF DISSOLVED OXYGEN IN LAKE ERIE 103
Abstract 103
Introduction 104
Data and methods of analysis 107
Clustering method 107
Application of clustering method 111
Depletion rate calculations and tests 112
Calculation of depletion rates 112
Detection of time effects in the oxygen concentration 115
Tests for constancy of depletion rate 115
Results and discussion 116

5. A STATISTICAL MODEL FOR DISSOLVED OXYGEN IN THE CENTRAL BASIN OF LAKE ERIE 131
Introduction 131
Model development 133
Prediction of the depletion rate and trend in the historical record 140
Prediction of anoxia in the hypolimnion 143
Conclusion 145

6. SPATIAL AND TEMPORAL DISTRIBUTION OF TOTAL AND FECAL COLIFORM CONCENTRATIONS 1966-1970 151
Introduction 151
Statistical methods 152
Empirical frequency distributions 153
Probability distributions for bacterial counts . . . 153
Fitting probability distributions and estimating parameters 155
Spatial zonation procedures 163
An example: analysis of 1968 total coliform data . 166
Inference about the ratio of bacterial counts . . . 171

Contents (Cont.)

Page

Analysis of total coliform data 173
Analysis of fecal coliform data 182
Summary and discussion 189

7. SAMPLING STRATEGY FOR FUTURE DATA COLLECTION 233
Introduction 233
Example to illustrate the principles of sampling . . . 235
Sampling strategy when the lake is divided into zones 238
Applications 240
Generalization 242
Sampling strategy when the concentration is a continuous function 244
Applications 245

REFERENCES 252

APPENDIX. Examination of linear regression models by residual analysis 265

Tables

Page

Table 1. Data on the Great Lakes 21

Table 2. Oligotrophic versus eutrophic 21

Table 3. Average Lake Erie Central Basin hypolimnion characteristics for 1973 to 1980 22

Table 4. Dissolved oxygen characteristics of the Central Basin of Lake Erie, 1930 to 1980 22

Table 5. The values of EC for the geographic zones of Lake Erie 61

Table 6. A summary of steps used in modelling the water level 62

Table 7. Central Basin air temperature: monthly means . . . 62

Table 8. Residuals and cumulative sum residuals for the Central Basin air temperature 63

Table 9. A summary of the air temperature pattern for the Central Basin, 1967 to 1978 63

Table 10. Results of fitting model 3.4 to the surface temperature data of Lake Erie, 1967 to 1978 64

Table 11. The values of F-statistic for testing the equality of yearly water temperature cycle 65

Table 12. The mean of the hypolimnion temperature, the temperature at Julian day 200 and the slope and intercept of the regression line, 1967 to 1980 65

Table 13. The estimates of the parameters of model 3.4 for chlorophyll a (Western Basin) 65

Table 14. The values of F_1, F_2 and F_0 for chlorophyll a (Western Basin) 66

Table 15. The estimates of the parameters of model 3.6 for chlorophyll a (Western Basin) 66

Table 16. The estimates of the parameters of model 3.4 for chlorophyll a (Central Basin) 66

Table 17. The values of F_1, F_2 and F_0 for chlorophyll a (Central Basin) 66

Tables (Cont.)

Page

Table 18. The estimates of the parameters of model 3.6 for chlorophyll a (Central Basin) 67

Table 19. The estimates of the parameters of model 3.4 for chlorophyll a (Eastern Basin) 67

Table 20. The values of F_1, F_2 and F_0 for chlorophyll a (Eastern Basin) 67

Table 21. The estimates of the parameters of model 3.6 for chlorophyll a (Eastern Basin) 67

Table 22. The estimates of the parameters of model 3.4 for TP (Western Basin) 68

Table 23. The values of F_1, F_2 and F_0 for TP (Western Basin) . 68

Table 24. The estimates of the parameters of model 3.6 for TP (Western Basin) 68

Table 25. The estimates of the parameters of model 3.4 for TP (Central Basin) 69

Table 26. The values of F_1, F_2 and F_0 for TP (Central Basin) 69

Table 27. The estimates of the parameters of model 3.6 for TP (Central Basin) 69

Table 28. The estimates of the parameters of model 3.4 for TP (Eastern Basin) 70

Table 29. The values of F_1, F_2 and F_0 for TP (Eastern Basin) 70

Table 30. The estimates of the parameters of model 3.6 for TP (Eastern Basin) 70

Table 31. The values of F_1, F_2 and F_0 for SRP (three basins) 70

Table 32. The estimates of the parameters of model 3.6 for SRP (three basins) 71

Table 33. The estimates of α_t and R_t for NO_3NO_2-N (three basins) 71

Table 34. The values of F_2 and F_0 for NO_3NO_2-N (three basins) 71

Table 35. The estimates of α_t and R_t for NO_3NO_2-N (three basins) 72

Tables (Cont.)

Page

Table 36. The values of F_1, F_2 and F_0 for SRP (three basins) 72

Table 37. The estimates of the parameters of model 3.6 for SRP (three basins) 72

Table 38. Results of fitting model 3.4 to log Secchi disk depth of Lake Erie, 1967 to 1978 73

Table 39. The values of F-statistic for testing the equality of the yearly Secchi disk seasonal cycle 74

Table 40. Results of fitting model 3.4 to log turbidity data of Lake Erie, 1967 to 1972 75

Table 41. The values of F-statistic for testing the equality of the yearly turbidity seasonal cycle 76

Table 42. The ratio of ammonia to nitrite and nitrate 99

Table 43. Summary of dissolved oxygen concentrations and depths for the groups restricted to common stations within a year, Central Basin, 1967 to 1979 121

Table 44. Summary of dissolved oxygen concentrations and depths for the groups restricted to common stations within a year, Eastern Basin, 1967 to 1980 122

Table 45. Summary of simple linear regression parameters and precision indicators from regression analysis of dissolved oxygen concentrations, Central and Eastern basins 123

Table 46. The estimates of the slopes and intercepts using weighted regression 124

Table 47. The estimates of the slopes and intercepts using empirical weighted regression 124

Table 48. Results of testing the equality of the slopes and the equality of the intercepts for model 4.1 125

Table 49. Estimate of the parameters of model 5.2 146

Table 50. Summary of the number of stations where total coliforms, fecal coliforms or both were measured for all available surface data 196

Tables (Cont.)

Page

Table 51. Observed frequency distribution, $f(r)$, of total coliform concentrations, r, for cruise 111 of 1966, 102 of 1968, and 107 of 1970 197

Table 52. Goodness of fit tests of the negative binomial distribution to the 1968 total coliform data using both the method of moments and maximum likelihood estimates for p and k 198

Table 53. Comparison of grouping procedures using total coliform data 199

Table 54a. Comparison of the fit of the negative binomial distribution to some of the 1967 total coliform data using all of the data for the cruise and a subset with high values removed 200

Table 54b. Comparison of the fit of the negative binomial distribution to some of the 1967 total coliform data using a subset with high values removed 200

Table 55. Test of fit of the negative binomial distribution to the total coliform data for 1967, 1968 and 1969 . . 201

Table 56. Estimates of the parameters of the negative binomial distribution for the total coliform data of 1967, 1968 and 1969 202

Table 57. Test of fit of lognormal distribution to total coliform data of cruise 111 of 1966 and cruise 107 of 1967 203

Table 58. Zones 1 to 5 for total coliforms by cruise from 1966 to 1970 204

Table 59. Total coliform concentrations and location of the stations in zone 5 205

Table 60. Observed frequency distributions for fecal coliform data, 1967 to 1975 206

Table 61. Conditional frequency distribution $f(R_f/R_t)$ for fecal coliforms, R_f, given total coliforms, R_t, based on 1967 and 1968 data 207

Table 62. Summary of fecal coliform concentrations at stations for which both total and fecal coliforms were measured 208

Tables (Cont.)

Page

Table 63. Summary of total coliform concentrations at stations for which both total and fecal coliforms were measured 209

Table 64. Ratio of fecal coliform concentration to total coliform concentration 210

Table 65. Turbidity and total coliforms by zone using only stations at which both turbidity and total coliforms were measured 211

Table 66. Mean, weighted mean, variance, efficiency, and the number of stations for each cruise 247

Table 67. Optimal allocation of sampling stations to Lake Erie 247

Table 68. Relative deviation and the corresponding sample size which is capable of detecting differences at 5% significance level 248

Illustrations

Page

Figure 1. The Great Lakes . . . 23

Figure 2. Lake Erie bathymetry . . . 24

Figure 3. Phosphorus concentration level of Lake Erie, 1970-1980 . . . 25

Figure 4. Mean depletion rates for dissolved oxygen during summer in bottom-water of central Lake Erie . . . 26

Figure 5. Historic oxygen depletion rates in the hypolimnion of central Lake Erie normalized to thickness conditions of 1970 . . . 26

Figure 6. The boundaries of the Western, Central and Eastern basins of Lake Erie . . . 77

Figure 7. The values of EC for each month using temperature (1968-1976) . . . 77

Figure 8. Annual mean water level for the Central Basin . . . 78

Figure 9. Five-year moving median . . . 78

Figure 10. Residual sum of squares against years . . . 78

Figure 11. The observed water level and the estimated water level against years . . . 79

Figure 12. The standardized residuals and orthogonalized residuals against years . . . 79

Figure 13. Cumulative standardized residuals against years . . . 80

Figure 14. Cumulative squared residuals and their expected values . . . 80

Figure 15. The values of $\hat{\alpha}_0$ against years . . . 81

Figure 16. The values of $\hat{\alpha}_1$ against years . . . 81

Figure 17. The values of $\hat{\alpha}_2$ against years . . . 82

Figure 18. The observed and estimated water level data using model 3.3 . . . 82

Figure 19. Spring, summer, fall and winter mean water level . . . 83

Illustrations (Cont.)

Page

Figure 20. Mean yearly Niagara River flows at Queenston . . . 83

Figure 21. The relationship between the Niagara River flow and Lake Erie water level 84

Figure 22a. Monthly air temperature means for the years 1967-1978 (Central Basin) 85

Figure 22b. Plots of the CUSUM for the air temperature for the years 1967-1978 (Central Basin) 85

Figure 23. The seasonal temperature cycles for the Western, Central and Eastern basins 86

Figure 24. The mean of surface water temperature for the Western Basin against that for the Central and Eastern basins and for the warming and cooling periods 86

Figure 25. The mean surface water temperatures and their estimated values from model 3.4 87

Figure 26a. Mean residual temperature for the Western Basin . . 87

Figure 26b. Mean residual temperature for the Central Basin . . 88

Figure 26c. Mean residual temperature for the Eastern Basin . 88

Figure 27. Temperature depth profile (July 29 to August 2, 1968) 89

Figure 28. Temperature depth profile for the Central Basin, 1967 90

Figure 29. Temperature depth profile for the Central Basin, 1978 91

Figure 30. Mean temperature values for the epilimnion and hypolimnion for the Central Basin during 1967 to 1978 and the fitted regression equations for the hypolimnion data 92

Figure 31. The mean uncorrected chlorophyll a against time . . 92

Figure 32. The estimates of the year effects against years (chlorophyll a) 93

Illustrations (Cont.)

Page

Figure 33. Mean TP against Julian days for the Western, Central and Eastern Basins 93

Figure 34. The estimates of the year effects against years (TP) 94

Figure 35. The plot of the year effects for chlorophyll a against those for TP 94

Figure 36. Time trend for SRP 94

Figure 37. Observed and estimated values for NO_3NO_2-N in the Western Basin 95

Figure 38a. Observed and estimated values for NO_3NO_2-N in the Western Basin for 1967, 1968 and 1970 95

Figure 38b. Observed and estimated values for NO_3NO_2-N in the Western Basin for 1971 96

Figure 38c. Observed and estimated values for NO_3NO_2-N in the Western Basin for 1978 96

Figure 39a. Observed and estimated values for NO_3NO_2-N in the Central Basin for 1967, 1968 and 1970 97

Figure 39b. Observed and estimated values for NO_3NO_2-N in the Central Basin for 1971 97

Figure 39c. Observed and estimated values for NO_3NO_2-N in the Central Basin for 1978 98

Figure 40. Observed and estimated values of NO_3NO_2-N in the Eastern Basin 98

Figure 41. Observed and estimated NH_3 values in the Western Basin 99

Figure 42. Observed and estimated NH_3 values in the Central Basin 99

Figure 43. Observed and estimated NH_3 values in the Eastern Basin 100

Figure 44. The estimated log Secchi disk value for each basin against years 101

Illustrations (Cont.)

Page

Figure 45. The observed Secchi disk depths and their estimated values from model 3.4 101

Figure 46. Average residual for Secchi disk 102

Figure 47. Relationship between chlorophyll a year effect and TP year effect 102

Figure 48. Flow diagram of the clustering algorithm 126

Figure 49. Example of computer plots produced by the clustering program for cruise 7022106 127

Figure 50. Standard deviation of the hypolimnetic oxygen concentration for clusters in the Eastern Basin plotted against Julian days 127

Figure 51. Plots of residuals against sequential order of the observations of the Central Basin: (a) standardized and orthogonalized residuals; (b) cumulative residuals; and (c) squared orthogonalized residuals 128

Figure 52. Plots of residuals against sequential order of the observations for the Eastern Basin: (a) standardized and orthogonalized residuals; (b) cumulative residuals; and (c) squared orthogonalized residuals 129

Figure 53. Dissolved oxygen depletion rates against year with fitted curves for the Central and Eastern basins . 130

Figure 54. Comparison of uncorrected dissolved oxygen depletion rates calculated by several authors 130

Figure 55. The estimated intercept of the dissolved oxygen regression line against Lake Erie water level . . . 147

Figure 56. The estimated depletion rate against Lake Erie water level 147

Figure 57. The estimated depletion rate from model 5.2 against Lake Erie water level 147

Figure 58. The plot of observed oxygen concentrations and their estimated values from model 5.10 against time in Julian days 148

Illustrations (Cont.)

Page

Figure 59. The plot of depletion rate against water level for different levels of TP 148

Figure 60. The 95% confidence interval for TP for the years 1948 to 1962 149

Figure 61. The predicted O_2 concentration from model 5.10 against Julian days for different temperature, water level and TP 150

Figure 62. The estimated probability of anoxia for a stratification period of 100 days and different levels for TP and the water level 150

Figure 63. The observed frequency distribution and the fitted Poisson and negative binomial frequency distributions, 1968 total coliform data 212

Figure 64. Contours of constant relative likelihood $R(p,k) = c$, $c = 0.01, 0.05, 0.50, 0.90, 0.99$, for p,k parameters of the negative bionomial distribution and total coliform data for cruises 102, 108, 109 and 112 of 1968 213

Figure 65a. Relative maximum likelihood functions of p, $R_m(p)$, for p a parameter of the negative binomial distribution and total coliform for cruises 102, 108, 109 and 112 of 1968 214

Figure 65b. Relative maximum likelihood functions for k, $R_m(k)$, for k a parameter of the negative binomial distribution and total coliform for cruises 102, 108, 109 and 112 of 1968 214

Figure 66. Probability distributions corresponding to groups determined for cruise 102, 1968 215

Figure 67. Observed frequency distributions for total coliform concentration (R) or $\log_e R$ shown on two scales . . 216

Figure 68. Mean total coliform concentrations for zones 1 to 4, zone 1 and zone 2, and maximum likelihood estimates of the parameters of the negative binomial distribution by cruise 218

Figure 69. Zones of Lake Erie based on total coliform concentrations, by cruise, from 1966 to 1970 219

Illustrations (Cont.)

Page

Figure 70. Examples of zone 2 characteristic of the spring and peak total coliform distributions 223

Figure 71. Zones 3, 4 and 5 based on the total coliform concentrations plotted together for all 1967 cruises . 223

Figure 72. Fecal against total coliform plotted on two ranges of total coliform concentrations, 1967 and 1968 cruises 224

Figure 73. Frequency distribution of fecal coliforms conditional on the total coliform concentration (R_t) shown for $R_t \leq 20$ and for three 1967 cruises and three 1968 cruises 225

Figure 74. Fecal coliform concentrations plotted against total coliform concentrations by zone and cruise for 1967 and 1968 226

Figure 75. Mean fecal coliform concentration against mean total coliform concentration for zone 1 of the 1967 to 1969 cruises 227

Figure 76. Ratio of fecal coliform concentration to total coliform concentration plotted against cruise date for zones 1, 2 and 3 using data from common stations 227

Figure 77. Location of stations at which fecal coliform concentrations were determined 228

Figure 78. Relative conditional likelihood function, $R_c(\rho)$, for the ratio of fecal to total coliform concentration for four cruises 230

Figure 79. Two examples of fecal coliform concentrations in the total coliform zones 230

Figure 80. Total coliform concentration against turbidity for stations in the Western Basin, 1968 cruises 231

Figure 81. Total coliform concentration against turbidity for 1968 cruises 231

Figure 82. Mean total coliform concentration against mean turbidity for zones 1 to 4 of the 1968 cruises using common stations in the zones determined from all the total coliform data 232

Illustrations (Cont.)

Page

Figure 83. The number of stations needed to detect a difference in the mean at the 5% significance level against relative deviations 249

Figure 84. The variance of the mean surface temperature as a function of the number of stations 250

Figure 85. The variance of chlorophyll a as a function of the number of stations 251

Figure 86. The percentage deviation for the mean chlorophyll a for different sample size 251

Abstract

This report presents a statistical evaluation of the water quality data available on Lake Erie from the Great Lakes surveillance program, 1966 to 1981. In Chapter 1, an introduction is given. Chapter 2 presents an up-to-date review of the early historical work on the eutrophication of Lake Erie. Chapter 3 discusses the problem of examining the data to determine the existence and nature of any trend that may be present in the values of the following limnological parameters: (1) the water level of Lake Erie for the years 1900 to 1979; (2) Niagara River flows for the period 1860 to 1975; (3) air and water temperature, Secchi disk depth, turbidity, total phosphorus, soluble reactive phosphorus, chlorophyll a, and nitrogen for the period 1967 to 1980. The analysis is performed for each of the Western, Central and Eastern basins of Lake Erie. Chapter 4 presents a clustering technique which automatically separates the data into the various naturally occurring thermal and spatial regimes. The resulting dissolved oxygen depletion rates, which are calculated on the basis of common stations in these clusters within a year, are found to be between those calculated by previous authors. Thus the cluster analysis presents a semi-objective and practicable alternative to the previous methods of data selection. Various types of regression analysis were applied to determine the hypolimnetic oxygen depletion rates for the years 1967 to 1980 in the Central and Eastern basins of Lake Erie. Statistical tests indicated that the depletion rate changed from year to year. The problem of explaining the differences in the depletion rate from year to year is considered in Chapter 5. The model uses the lake water level, the hypolimnion temperature and yearly mean total phosphorus concentrations as explanatory variables. It enables the estimation of the probability of anoxia as a function of the three explanatory variables. Also, the use of the model for

setting regulations and for specifying standards is given. The characterization of the spatial and temporal variabilities of total coliform and fecal coliform concentrations in Lake Erie is determined using the available data for 1966 to 1970. Furthermore, the relationship between turbidity and total coliform concentration is considered. Also in this chapter, some general statistical methodologies are given which are useful for the analysis of discrete data (i.e. counts) including other limnological variables such as phytoplankton, zooplankton and fish. In Chapter 7, the question is investigated concerning how to use past information about the spatial and temporal variabilities of limnological data to plan a strategy for future data collection when the aim is to estimate the areal weighted mean value of a single limnological variable. This statistical approach is applied to coliform counts, temperature and chlorophyll a.

Résumé

Le présent rapport est une évaluation statistique des données relatives à la qualité de l'eau du lac Érié obtenues de 1966 à 1981 dans le cadre du programme de surveillance des Grands lacs. On trouvera un résumé au premier chapitre et, au chapitre 2, une mise à jour du compte rendu des premiers travaux portant sur l'eutrophisation du lac Érié. Le troisième chapitre traite du problème de l'examen des données pour déterminer l'existence et la nature de toute tendance que pourraient présenter les valeurs des paramètres limnologiques suivants : (1) le niveau de l'eau du lac Érié entre 1900 et 1979; (2) le débit de la rivière Niagara entre 1860 et 1975; et (3) la température de l'air et de l'eau, la profondeur au disque de Secchi, la turbidité et les teneurs en phosphore total, en phosphore réactif soluble, en chlorophylle a et en azote pendant la période allant de 1967 à 1980. Une analyse est effectuée pour chacun des bassins, ouest, central et est, du lac Érié. On présente au chapitre 4 une technique d'analyse par grappes qui permet de regrouper automatiquement les données selon les divers régimes thermiques et spatiaux naturels. Il s'avère que les taux d'utilisation de l'oxygène dissous calculés à partir des stations visitées à toutes les sorties dans ces grappes pendant une année sont compris dans ceux calculés par les chercheurs précédents. L'analyse par grappes est donc une solution de remplacement semi-objective et pratique pour les méthodes antérieures de choix de données. Divers types d'analyse par régression ont été utilisés pour déterminer les taux d'épuisement de l'oxygène hypolimnétique au cours des années 1967 à 1980 dans les bassins du centre et de l'est du lac Érié. Des tests statistiques ont montré que le taux d'épuisement a varié d'une année à l'autre. On traite, au chapitre 5, du problème que pose l'explication de ces différences annuelles. Le modèle utilisé fait appel au niveau de

l'eau, à la température de l'hypolimnion et aux concentrations moyennes de phosphore total annuelles comme variables explicatives. Le modèle permet d'estimer la probabilité d'anoxie en fonction de ces trois variables. On traite aussi de l'utilisation du modèle pour l'élaboration de réglementations et la détermination de normes. Les données obtenues entre 1966 et 1970 ont servi à caractériser les variabilités spatiales et temporelles des teneurs de coliformes totaux et fécaux du lac Érié. De plus, on s'intéresse à la relation entre la turbidité et la teneur de coliformes totaux. On présente aussi, au même chapitre, certaines méthodes statistiques générales utiles à l'analyse de données discrètes (dénombrements), y compris d'autres variables limnologiques comme le phytoplancton, le zooplancton et le poisson. On s'intéresse, au chapitre 7, à la façon d'utiliser les vieilles informations sur la variabilité spatiale et temporelle des données limnologiques dans le but de planifier une stratégie pour la cueillette des données quand on désire estimer la valeur moyenne, pondérée en fonction de la superficie, d'une seule variable limnologique. Cette approche statistique est appliquée aux dénombrements des coliformes, à la température et aux teneurs de chlorophylle a.

Acknowledgments

The contributors to this report would like to express their appreciation to all who helped directly or indirectly in completing this task. Special thanks are due to K. Rodgers, F. Elder, J. Barica, R.A. Vollenweider, B.J. Dutka, P. Hamblin, C.R. Murthy, D.C. Lam, and D. Warry for their encouragement, support and discussion.

Our thanks are due to M. Fellows and A. [illegible] for programming assistance. We would also like to thank C.L. Minnie for the excellent typing of the text, and W.D. Finn and his associates for the excellent job of drafting the figures.

We greatly acknowledge the financial support received under the Great Lakes Water Quality Agreement.

CHAPTER 1

Introduction

The purpose of this report is to use statistical techniques to summarize the historical information available on Lake Erie from the Great Lakes surveillance program, 1966 to 1981. This program was designed and implemented by the Canadian government to provide detailed and semi-systematic information about the spatial and temporal changes in the water quality of the Great Lakes. Prior to this program there was sparse information on the biological, chemical and physical processes occurring in the lakes. Certainly, the collected data have advanced our understanding of the dynamics of most limnological characteristics and of the interactions between them. Moreover, the collected data have allowed us to re-examine and modify our concepts and to test different models and hypotheses regarding the factors controlling water quality. However, only a small portion of the information available in the surveillance data has been examined scientifically. This is because most studies were either restricted to investigating the characteristics of a small number of limnological variables, examining only the data collected during a number of cruises, or using inadequate methods of data analysis. Comprehensive and integrated statistical evaluation of all the collected data, as opposed to isolated and limited studies, would ensure the utilization of all the information available from the surveillance program. Such an analysis would permit the isolation of the influence of different sources of variability (spatial and temporal) and the study of the association between different variables through the use of empirical models. Furthermore, as the surveillance program is more than 15 years old, it is time to examine its capabilities and the usefulness of its information. Modifications could then be made to the program where appropriate.

Specifically, this report discusses the following:

1. Extraction of the information available in the surveillance data and its utilization for (a) examining the changes in the quality of water (i.e., determining the time trend), (b) examining the spatial variability and isolating regions in the Great Lakes with waters of low quality, (c) developing empirical models and studying the association between different limnological variables.

2. Development of a strategy for future data collection.

3. Development of methodologies that could be applied to similar environmental problems such as studying the spatial and temporal variabilities of parameters measured to study the effect of acid rain.

CHAPTER 2

Lake Erie Review

by R.E. Kwiatkowski

The International Lake Erie Water Pollution Board and the International Lake Ontario-St. Lawrence Water Pollution Board (1969) have recommended that the governments of Canada and the United States agree to develop a joint program to control the pollution of the Great Lakes ecosystem. In the Great Lakes, three elements (phosphorus, nitrogen and silicon) have been found to limit aquatic plant productivity. The International Joint Commission (IJC) has given a great deal of attention to phosphorus controls because this nutrient can be the most readily removed from water bodies, on both a technical and an economical basis. In 1972, the Great Lakes Water Quality Agreement (72WQA) was signed between Canada and the United States. The Agreement recognized that eutrophication was a major problem in the lower Great Lakes (Erie and Ontario), and the IJC assigned to the Great Lakes Water Quality Board the responsibility of developing phosphorus loading objectives to alleviate the problem. In the fifth year of the Agreement, a comprehensive review of the established programs was done and a new Great Lakes Water Quality Agreement was signed in 1978 (78WQA), which contained total phosphorus loading objectives for each of the Great Lakes. In contrast with the approach used in the 72WQA, the 78WQA objectives were based on the resulting water quality corresponding to the phosphorus loads in each basin.

Restoration of the year-round aerobic conditions in the bottom waters of the Central Basin of Lake Erie was stated as a primary objective in Annex 2 of the 72WQA and Annex 3 of the 78WQA. To achieve this goal much scientific research over the past ten years has been done. A brief review of the early historical work concerning dissolved gases in lakes and of the factors controlling oxygen depletion in hypolimnetic waters together with a review of some of the major works on the extent of anoxia in Lake Erie are presented here. It is not the intent of this chapter to present a detailed or complete review on oxygen depletion in temperate lakes. The purpose is to give the reader a greater appreciation of the complex biological, chemical and physical interactions which take place in lakes, and of the wealth of information obtained concerning the eutrophication-oxygen depletion problem in the hypolimnetic waters of Lake Erie, the smallest of the Great Lakes (Table 1, Fig. 1).

Lake Erie is actually three lakes in one. Much of the lake is shallow, the small Western Basin is mostly less than 11 m in depth, while the large Central Basin has a maximum depth of about 25 m. The Eastern Basin is the deep portion of the lake with depths of 60 m (Fig. 2). Details on the bedrock geology of the three basins can be found in Sly (1976), and information on water transport between the basins is given by Simons (1976).

EARLY HISTORICAL WORK ON DISSOLVED GASES

The importance of O_2 depletion in hypolimnetic waters as a measure of biological activity was discussed by Hoppe-Seyler (1895). From the measurements of O_2 concentrations from only five depths (maximum depth 245 m) in Lake Constance, Hoppe-Seyler realized the importance of oxygen studies in understanding the biological processes taking place in the water, and the concept of oxygen-deficit was introduced to limnology. In 1911, Birge and Juday discussed the annual cycle of thermal stratification, and the utilization and production of O_2 and CO_2 by biological processes in lakes. Earlier, Birge (1906) had written that O_2 decreases in the hypolimnion were a function of four main factors: the quantity of decomposable material (autochthonous and allochthonous) contributed to it from the epilimnion; the volume of the hypolimnion (which in turn depends on the depth of the lake); the length of time that the bottom water was cut off from the epilimnion; and the temperature of the bottom water (through its effect on the rate of decomposition).

August Thienemann used oxygen depletion values as a quantitative measurement of the degree of eutrophy (eutrophic-rich in nutrients versus oligotrophic-poor in nutrients) in a lake. Thienemann (1915) determined five factors affecting the degree of oxygen decrease in the hypolimnion: (1) the season of the year; (2) the position of the lake (with respect to prevailing winds); (3) the magnitude of the volume constituting the hypolimnion, and the ratio between the volume of water above and below the thermocline; (4) the temperature of the hypolimnetic waters; and (5) the quantity of organic matter transported into the bottom waters. Thienemann (1928) later stressed factor (3) as extremely important, giving special significance to the ratio of the epilimnion volume to hypolimnion volume. Furthermore, Thienemann concluded that since the deep water of shallow lakes _ceteris paribus_ warms more rapidly than that of deep lakes, and as heat enhances decomposition, the difference in the heating of these two lakes owing to their different depths exerts an influence in the same direction as does the difference in the volume of their tropholytic layers.* Thienemann (1928) established the following morphologic characteristic for an oligotrophic lake. The volume of water in the hypolimnion is large relative to that in the epilimnion (Table 2), whereas for a eutrophic lake the volume of water in the hypolimnion is small relative to that in the epilimnion.

---

* Trophogenic - the superficial layer in which organic production takes place on the basis of light energy; tropholytic - the deep layer where organic dissimilation predominates because of light deficiency.

PHYSICAL, BIOLOGICAL AND CHEMICAL VARIABLES CONTROLLING D.O. DEPLETION

The thermocline acts as a barrier to the movement of dissolved gases from the epilimnion to the hypolimnion. Thus the formation of the annual temperature cycle is significant in the study of oxygen depletion. In temperate lakes during winter, ice covers the lake. A permanent stratification (winter stagnation) can be set up, with water temperatures of 0°C immediately below the ice and uniformly low temperatures at or slightly above 4°C in the deeper strata. After ice breakup, wind-generated currents result in the entire water mass being mixed (spring overturn) and the lake becomes isothermal and chemically homogeneous. After the spring equinox, heat energy is absorbed by the upper layers of the lake. However, absorption of increased solar radiation in spring has been shown to account for only a small portion of the total energy that distributes heat in a lake (Birge and Juday, 1921). The main energy source is the wind (Birge, 1916), which generates currents (the speed and direction of which are dependent on the strength and direction of the wind [Hutchinson, 1967]). When the surface water particles pushed by the wind reach the shore, they are deflected by the resistance of the colder and heavier deep water. As a result, a counter current is established just below the surface which leads to heat exchange (eddy diffusion). Due to density gradients established by the progressive accumulation of heat (water being the most dense at 4°C) a boundary layer (thermocline) is thus formed in sufficiently deep lakes between the totally intermixed surface layer (epilimnion) and the quiet water masses underlying it (hypolimnion) (Birge, 1910). Birge (1898) defined the thermocline as the region of rapid decrease in temperature in which the gradient was greater than 1°C per metre. Brönsted and Wesenberg-Lund (1912) redefined the term to mean the plane of maximum rate of decrease in temperature, while Hutchinson (1957) designated the whole region in which the temperature gradient was steep, from the upper plane of maximum curvature, termed "the knee" of the thermocline, to the lower plane of maximum (inverse) curvature, as the metalimnion. It should be pointed out that owing to the complicated interactions of solar radiation, wind and morphological features of a lake, lakes often exhibit individual temperature curves. As weather conditions change from year to year, a given lake may exhibit unique temperature curves each year (Table 3), with varied epilimnion depths.

When the loss of heat in a lake is greater than heat intake, thermal stratification breaks down. Radiation accounts for the greatest heat loss, with further losses through evaporation and through conduction to the air and the bottom sediments, which result in vertical convection currents. Also, the volume of water flowing through the lake, which in general carries away the topmost water strata, can be of substantial importance to the thermal economy (Ruttner, 1952). Eventually by late fall, isothermal conditions return and the lake again becomes totally mixed (fall overturn).

The importance of the thermocline to the biology of a lake was first described in a classic paper entitled "The Thermocline and its Biological Significance" (Birge, 1904). The fundamental process of life, carbon assimilation, occurs mainly but not exclusively in the upper waters of a lake (trophogenic layer). The rates and the amount of assimilation are governed by a complicated set of factors dependent on the species of phytoplankton present, solar radiation, nutrient availability, temperature and the interrelationships between phytoplankton requirements. A review of these complex interactions has been done by Hutchinson (1967) and Lund (1965) and they will not be discussed further. Ultimately, the organic matter formed settles out of the trophogenic layer (often equivalent in depth to the epilimnion layer [Thienemann, 1928]) and settles into the deep tropholytic layer. This movement of organic material results in a decrease in the dissolved oxygen content of the hypolimnetic waters owing to plant, animal and bacterial respiration (in the decomposition of the organic matter), as well as by purely chemical oxidation of the organic matter in solution. Since the thermocline acts as a barrier to the movement of dissolved oxygen from the epilimnion, an oxygen deficit is created in the hypolimnion. Total oxidation of the organic matter does not always occur in the hypolimnion (water oxygen demand). Large amounts of organic matter in eutrophic lakes settle to the bottom either to be oxidized (sediment oxygen demand) or to be stored. A microstratification of deoxygenated water forms between the bottom sediments and the overlying waters. This microlayer is gradually enlarged by eddy diffusion, and in highly eutrophic lakes, the entire hypolimnion can become anoxic. Anoxic conditions in the hypolimnion have obvious deleterious effects on its biological community (benthos, zooplankton, bacteria and cold water fisheries) and also can affect its chemistry. If the dissolved oxygen at the mud-water interface is significantly reduced, ferric and manganic hydroxides are solubilized. Reduction of ferric phosphate results in the release of phosphorus from the sediments and back into the water column (internal loading), resulting in further enrichment of the lake.

PRESENT STATE OF LAKE ERIE

Because of a serious decline in certain fish populations in Lake Erie, a limnological study was undertaken on Lake Erie in 1928 and 1929 (Fish, 1960). Sixty-two stations were sampled for a variety of physical, chemical and biological parameters on eight cruises between May and September. It was found that the mean hypolimnetic summer oxygen saturation was 83.3%. Burkholder (1960) concluded that the oxygen conservation in Lake Erie (for 1929) appeared to be due to the general oligotrophic character of the lake. Decomposition was reported as moderate and thus minimum oxygen levels, as exist in eutrophic lakes, were not found. By the late 1960's, anoxic conditions in the hypolimnion of Lake Erie became commonplace and Lake Erie was referred to as a dead lake.

Commercially valuable species such as blue pike, whitefish, lake herring, and sauger have either drastically declined or disappeared, having been replaced by such less desirable (commercially) forms as alewife, smelt and freshwater drum (Baldwin and Saalfield, 1962; Leach and Nepsy, 1976). After a review of the pertinent literature, Leach and Nepsy concluded that the species shifts were due to a variety of stresses, which in order of importance are intensive commercial fishing; nutrient loadings (resulting in an anoxic hypolimnion); introduction of exotic species; tributary and shoreline reconstruction; erosion and siltation; and the introduction of toxic materials.

Beeton (1965), in a study on the eutrophication of the St. Lawrence Great Lakes, concluded that man's activity had clearly accelerated the rate of eutrophication of Lake Erie. His conclusions were based on a trend analysis of chemical data from 1854 to 1960. Beeton (1965) observed large increases in chloride and sulphate concentrations, both of which are conspicuous in domestic and industrial wastes. The changes in concentrations of these parameters paralleled the population growth in the Lake Erie basin. Comparison of recent data from the Central Basin nearshore zone with the values of Beeton (1965) indicated that the 1978/79 concentrations were simlar to those of the late 1950's, and indeed, for calcium and chloride, were actually lower (Richards, 1981). Richards concluded that the water quality of Lake Erie was not deteriorating at the rate which typified the first half of the century.

A biological study of the benthic community of Lake Erie's Western Basin by Carr and Hiltunen (1965) indicated that more severe environmental conditions existed in 1961 than in 1930. The population of the burrowing mayfly, Hexagenia spp., was reduced from an average of $139/m^2$ in 1930 to less than $1/m^2$ in 1961.* Oligochaeta, a taxon typical of low oxygen conditions, had undergone a ninefold increase in numbers over the same time period ($677/m^2$ in 1930 to $5949/m^2$ in 1961; Carr and Hiltunen, 1965). As noted by the Phosphorus Management Strategies Task Force (1980), however, these major changes in the benthic community occurred during periods in which toxic materials such as DDT were first used extensively in the environment, and thus cause and effect relationships are difficult to establish.

The only long-term studies on the phytoplankton communities of Lake Erie were by Davies (1964, 1969). The average number of phytoplankton (cells/mL) showed an annual increase of 443 cells/mL/yr between 1929 and 1962, with a major increase occurring in the abundance of blue-green algae (a group commonly associated with eutrophication). It should, however, be noted that these data are from a water treatment plant at Cleveland, and therefore the changes in the phytoplankton reported by Davies (1964) are not indicative of offshore waters, but rather can be considered as representing the effects of localized inputs. A study of diatom frustules in cores from the

---

* In the 1980 Report on Great Lakes Water Quality, the IJC states that benthic studies in the Western Basin in 1979 have shown notable improvements in the benthic community. Hexagenia limbata reappeared for the first time since the early 1950's near the mouth of the Detroit River (IJC, 1980a).

Central Basin of Lake Erie by Harris and Vollenweider (1982) indicated that a major shift in the species composition of this taxa occurred around 1850 as a result of deforestation. A second minor shift (in relative abundance and not species) occurred in the mid-1900's and was probably due to increased phosphorus addition to Lake Erie via detergents. In conclusion, Harris and Vollenweider (1982) stated that the diatom composition has remained relatively constant for the last 100 years.

Excessive phosphorus levels have been identified as the most important factor in the accelerated eutrophication of the Great Lakes (Phosphorus Management Strategies Task Force, 1980). Numerous authors over the past 20 years have established phosphorus-chlorophyll-phytoplankton relationships in lakes (for a review see Nicholls and Dillon, 1978). From the relationships established it is apparent that increasing phosphorus loadings to Lake Erie via sewage discharge, detergents, agricultural runoff or the atmosphere will result in increased phosphorus concentrations in the lake. This theoretically will lead to increased algal biomass and subsequently to an increased areal extent of hypolimnetic anoxia during summer stratification.

In 1977, the IJC indicated that mean annual phosphorus concentrations had been significantly ($P \leq 0.05$) increasing in the Central Basin since 1973 (IJC, 1977). In 1979, however, from annual open lake monitoring of Lake Erie, the IJC concluded that total phosphorus concentrations had fluctuated only slightly since 1970 (IJC, 1979b). In 1980, the IJC concluded that no significant change in the phosphorus concentration level (Fig. 3) had occurred over the last decade (IJC, 1980b). The non-conservative nature of phosphorus, its large year to year variability, internal recycling, and its extensive interaction with biological populations made it very difficult to trace phosphorus concentration changes in the Great Lakes (Phosphorus Management Strategies Task Force, 1980).

Since 1973, the IJC has calculated total phosphorus loadings to the Great Lakes via monitored and unmonitored tributaries, direct inputs from municipal and industrial sources, connecting channels and the atmosphere. Nonpoint source estimates commenced in 1976. Unfortunately, no coordinated collection of phosphorus loading data and no scientifically sound methodology for calculating phosphorus loading exist (Phosphorus Management Strategies Task Force, 1980). This was demonstrated by the Task Group's example of the different total phosphorus loadings calculated for Lake Erie in 1976. The Task Group III (a U.S.-Canada review group established during the development of the 1978 Great Lakes Water Quality Agreement) calculated a loading of 19 500 tons/year based on the Lake Erie Wastewater Management Study by the U.S. Army Corps of Engineers. The Water Quality Board estimated the load to be 15 500 t/yr from the same data set, while PLUARG calculated a loading of 17 450 t/yr. Even with a single agency calculating the loadings, there is no guarantee of standardization. Different methods of calculating point source phosphorus loadings for the years 1972 to 1977 were used by the Water Quality Board (Zar, 1980). Zar concluded that greater efforts were needed to assess the quality and the statistical accuracy of the data. None of the loading estimates for calculating total phosphorus loading to Lake Erie took phosphorus regeneration from the bottom sediments into account, even though this can be equivalent to 111% of the external loading during anoxic conditions (Burns and Ross, 1972a). Even during oxic conditions phosphorus regeneration is important. Internal oxic and anoxic regeneration for a two-month summer period in Lake Erie was found to equal 137% of the external loading. Controversy also exists over whether or not control programs should be based on total phosphorus or on algal available phosphorus entering the water body (Lee et al., 1980). Since there are many different forms of phosphorus and their availability varies widely, it is almost impossible even to develop routine analytical procedures to quantify each biologically available form which may be present. Present eutrophication control programs are thus based on the concept of controlling all forms of phosphorus, irrespective of whether the phosphorus is in a form which can support algal growth or not (Lee et al., 1980).

Because of the growing concern over the increasing area of anoxia in the Central Basin of Lake Erie, an intensive Canada-United States study of the lake was undertaken in 1970 (Project Hypo). Twenty-five water sampling stations were sampled on ten surveys in the Central Basin, with an additional 16 stations established to provide bathythermograph records. Five of the 25 stations were termed "major stations" and were sampled intensively for the four disciplines (chemical, biological, bacteriological and physical) deemed most important to the study of oxygen depletion (Burns and Ross, 1972a). The conclusion drawn by Burns and Ross (1972b) and from the 1970 study was that "phosphorus input to Lake Erie must be reduced immediately; if this is done, a quick improvement in the condition of the lake can be expected; if it is not done, the rate of deterioration of the lake will be much greater than it has been in recent years." Anoxic conditions in the Central Basin hypolimnion in 1970 were unsuitable for fish life and this resulted in the regeneration of large amounts of phosphorus from the sediments (Burns, 1976b).

Beeton (1965) found low dissolved oxygen concentrations in 1959 and 1960, but stated that scattered observations of low oxygen had been reported for the past 33 years on Lake Erie. Dambach (1969) reported that a dissolved oxygen value of 0.8 mg/L was measured in

Lake Erie in 1929. However, from a study of the historical records of dissolved oxygen in the hypolimnion of Lake Erie, Dobson and Gilbertson (1971) concluded that the 1970 depletion rate was more than double the rate estimated for 1929. They further stated that central Lake Erie had become mesotrophic around 1940 and has been in the process of becoming eutrophic (as of 1970). The increase in deoxygenation was attributed to increases in phytoplankton production caused by increased nutrient inputs (Fig. 4). Charlton (1980a) reviewed the work of Dobson and Gilbertson and indicated that there was no long-term trend to increasing oxygen depletion. According to Charlton, differences in depletion rates that did occur between years were mostly related to variations in hypolimnetic thickness (Fig. 5). In fact, present day oxygen depletion rates, when corrected for the relatively high temperatures in Lake Erie's hypolimnion, are indicative of mesotrophic lakes (Charlton, 1980a). In a later paper, Charlton (1980b) stated that the use of oxygen concentration to compare lake productivity was not justified without reference to hypolimnion thickness and temperature (see conclusions of Birge and Juday [1911] and Thienemann [1928] given previously). Reassessment of the data by Rosa and Burns (1981) utilizing correction factors for hypolimnion thickness, vertical mixing, and thermochemical effects on a representative area in the Central Basin indicated that a significant increase in oxygen depletion rates from 1929 to 1980 had occurred.

A review of the annual Great Lakes Water Quality reports between 1973 and 1978 by the International Joint Commission leads to a somewhat confusing picture. It was clearly evident to the IJC (1975) that there had been a doubling of the anoxic hypolimnion between 1930 and the mid-1960's (Table 4). In 1975, a dramatic reduction in the anoxia was reported, a result of the formation of a deep hypolimnion caused by meteorological conditions at the time of thermocline formation (IJC, 1976a). A large anoxic area occurred again in 1976, a result of a relatively thin hypolimnion which was more easily depleted than in 1975 (IJC, 1977). The Commission (IJC, 1977) concluded that spring wind conditions and the time of initial stratification, which determine hypolimnion volume, were important factors in influencing the extent of anoxic conditions. In 1977, there was an apparent decrease in the area of anoxia; the decrease, however, was not due to changes in hypolimnion thickness, phosphorus concentrations, or algal production, but rather to a change in the definition of the term anoxic. Prior to 1977, a region with oxygen concentrations less than 1.0 mg/L was considered anoxic, whereas in 1977, only regions with oxygen concentrations less than 0.5 mg/L were considered anoxic (IJC, 1978a, 1978b).* The IJC concluded that the area of anoxia was not an accurate measure of trophic status for Lake Erie (IJC, 1979b).

A review of the volumetric oxygen demand (mg O_2/L/d) in the Central Basin by the IJC (1976a) showed a marked increase between 1930 and 1970. The oxygen demand rate had more than doubled over the 40-year period (Table 4). An intensive surveillance study was conducted on Lake Erie in 1978 to 1979 as required by the Great Lakes International Surveillance Plan (GLISP) to gather more information on these apparent trends. The highlights drawn from the first year of the two-year intensive program were that the areal extent of anoxia in Lake Erie was not an accurate measure of the lake's trophic status (IJC, 1979b) because of the effects of water level and meteorological conditions. Furthermore, the volumetric oxygen depletion rate in the Central Basin had not changed since 1970 (Table 4), confirming the conclusion that the overall eutrophic status of the lake had not changed (IJC, 1979a). The importance of volumetric oxygen depletion rates and of their relationship to environmental management strategies in the control of eutrophication was difficult to assess (IJC, 1979b). The IJC (1979a) concluded that the low dissolved oxygen in Lake Erie's Central Basin was either a result of increased phosphorus discharges (cultural eutrophication) or an ongoing situation that had not appreciably changed over the past 20 or 30 years.

---

* In Appendix B (IJC, 1978b), the area of anoxia is defined as the area with oxygen concentrations $\leqslant$0.5 mg/L. This is visually displayed in Figure 2.2-8. At the top of Table 2.2-2 (IJC, 1978b), however, the definition of anoxia is given as the area with oxygen concentrations $\leqslant$1.0 mg/L, as had been reported in previous IJC reports. The 1977 areal extent of anoxia calculated from $\leqslant$0.5 mg/L is included in the table, while in the text reference is made to the fact that the areal extent of anoxia in the hypolimnion during 1977 was less than that reported in 1976. There is no mention of the fact that the 1975 value and all previous values for areal extent are at the 1.0 mg/L level.

A Phosphorus Management Strategies Task Force was asked to review and comment on the different scientific opinions. From its studies, it concluded that the restoration of aerobic conditions to the Central Basin was a prime concern, and to restore year-round aerobic conditions, phosphorus concentrations must be controlled (Thomas et al., 1980). Attempts were made, via a combination of objective analysis and expert opinion, to establish the upper limits for nutrient concentrations which would enable a desired ecosystem to exist in each of the Great Lakes. Five different mathematical models were put forward to assist the Task Group in estimating lake responses to changes in phosphorus loadings (Bierman, 1980; Chapra, 1980; DiToro, 1980; Thomann and Segna, 1980; Vollenweider et al., 1980). However, only three of the modelling efforts (Chapra, 1980; DiToro, 1980; Vollenweider et al., 1980) had components within the model dealing with the relationship between phosphorus loads and hypolimnetic oxygen depletion.

As pointed out by Bierman (1980), none of the models that were reviewed had been tested for phosphorus loads other than the present ones, since there had not been a significant change in phosphorus loads during the period for which comprehensive in-lake data existed (1967 to 1980). Bierman further stated that the predictions given by the models were strictly best estimates and not absolute guarantees of future conditions. For a given input load, calculated phosphorus concentrations were within a range of about 5% to 25%. However, calculated chlorophyll a responses for Lake Erie differed radically between models (Phosphorus Management Strategies Task Force, 1980). Since oxygen depletion is a direct result of phytoplankton (chlorophyll a) concentration and not of phosphorus concentration, the proposed models can only give limited insight with respect to the dissolved oxygen content in the hypolimnetic waters of Lake Erie. A recently proposed model by Vollenweider and Janus (1981) predicted that a rapid decrease in oxygen depletion rates in Lake Erie would not occur until yearly average chlorophyll a concentrations fell below 2 mg/m^3. The model predicted that this would not occur until total phosphorus discharge to Lake Erie was less than 8000 tons/year.

Birge and Juday (1911) and Thienemann (1928) originally noted that the hypolimnion thickness is of primary importance in the calculation of oxygen depletion rates. This fact has more recently been addressed by Charlton (1979, 1980a, 1980b). As previously described, hypolimnion thickness is determined by water levels and meteorological conditions. Thus, ultimately the prediction of dissolved oxygen concentrations is dependent on the prediction of weather conditions, which at best is only valid for short-term projections. The models proposed are thus reduced to two possible alternatives. The model can be used a posteriori with knowledge of the weather conditions, or it can be based on a normally occurring condition (setting hypolimnion thickness to some arbitrary value). Probably the best comment on the efforts of mathematics to model the area of anoxia in Lake Erie mathematically was made by Thienemann in 1928: "To some extent we violate nature when we intend to formulate by numerical means complex, partly biological, natural phenomena like the oxygen condition existing in lakes."

CONCLUSIONS

Over the last ten years much has been published concerning the Lake Erie eutrophication problem (Limnological Survey of Eastern and Central Lake Erie, 1928 to 1929; Fish, 1960; Project Hypo, Burns and Ross, 1972a, 1972b; J. Fish Res. Board Can. Special Issue Vol. 33, "Lake Erie in the Early Seventies"). It is universally accepted that Lake Erie is eutrophic. Yet controversy still continues over whether or not statistically significant changes in water quality can be detected. The controversy exists because such a large amount of money is spent on Great Lakes Surveillance programs to determine the effect of remedial actions. In 1978 (IJC, 1978a), expenditures for the Great Lakes Surveillance plan were calculated to be 9.4 million dollars, of which 4.8 million was spent on the Lake Erie intensive study (1978 and 1979 are classified as intensive years for Lake Erie in the IJC's GLISP nine-year cycle). The Great Lakes Surveillance budget is split fifty-fifty between Canada and the United States. To put the cost of the surveillance program into perspective, it is less than 1% of the total amount expended annually on the implementation of pollution abatement programs in the Great Lakes basin.

Some of the past work on Lake Erie has been presented in this chapter. From this review, it is apparent that cause-effect relationships are difficult to establish in Lake Erie, as are statistically significant trends. Historical changes in the biological community may be due to eutrophication or to the introduction of toxic compounds (benthos; Phosphorus Management Strategies Task Force, 1980), exploitation (fish; Leach and Nepsy, 1976) or indicative of localized effects (phytoplankton; Davies, 1964). Changes in total phosphorus loadings are plagued by inconsistent methods of calculation and poor data (Zar, 1980), while phosphorus concentration levels in the lake are so variable between years that useful analysis does not seem possible (Phosphorus Management Strategies Task Force, 1980). Oxygen depletion rates in Lake Erie are on an increase (Dobson and Gilbertson, 1971) due to cultural eutrophication, or there have not been any appreciable differences in oxygen depletion rates over the past 30 or 40 years (Charlton, 1979, 1980a), and any changes in depletion rates are a result of changing hypolimnion thickness. Beeton (1965) reported that changes in water chemistry of Lake Erie were a result of man's influence, but found that reports of low dissolved oxygen concentrations had been made over the last 33 years. Finally, modelling efforts are reported as best estimates and no guarantees to restoration of aerobic conditions can be given, even if loading rates meet the newly set objectives of the 78WQA (Bierman, 1980; Thomas et al., 1980).

It should be pointed out that the IJC is not directly responsible for the research programs on the Great Lakes. This function is performed by scientists from various government agencies and the university community. The IJC's role is restricted to advice, coordination and the reporting of the findings. Thus the study of the processes within the lake that affect the trophic status of Lake Erie must ultimately be the responsibility of the research community.

There is an apparent need for a sound statistical review of the historical data gathered on Lake Erie for a variety of limnological parameters. This should be done not only to establish where any trends exist but also to determine whether there are any weaknesses in the data sets. At an annual cost of nearly ten million dollars for the Great Lakes surveillance program, it is mandatory that the most efficient cost-information rates be established.

Table 1. Data on the Great Lakes[1]

	Superior	Michigan	Huron	Erie	Ontario
Total basin area (km^2)[2]	127 700	118 100	133 900	58 800	70 700
Surface area (km^2)	83 300	57 850	59 510	25 820	18 760
Lake volume (km^3)	12 000	5 760	4 600	540	1 720
Average depth (m)	145	99	76	21	91
Maximum depth (m)	307	265	223	60	225
A. depth/max. depth	0.47	0.27	0.34	0.33	0.40
Retention time (yr)[3]	185	104	27	2.7	7.8

[1] Hutchinson (1957).

[2] Great Lakes Basin Commission. 1976. Limnology of Lakes and Embayments Great Lakes Basin Framework Study, Appendix No. 4, NOAA.

[3] Rousmaniere. 1979. The Enduring Great Lakes. New York: W.W. Norton and Co.

Table 2. Oligotrophic versus Eutrophic

Parameter[1]	Thienemann's oligotrophic/ eutrophic	Lake Superior[2] July 27 - Aug. 7/73	Lake Erie[2] Aug. 28 - Aug. 31/73	Central Basin[2] Aug. 28 - Aug. 31/73
$\bar{d}$	20	145	21	16
% Volume E	50	7	48	52
O_2 H/E	1.0	1.10	0.78	0.67
Sat. O_2 bottom	50	102.8	49.1	33.4
ΔE mL O_2/1000 mL	-1.0	0.85	0.11	0.10
ΔH mL O_2/1000 mL	-1.0	0.50	-2.20	-2.68
ΔH + E mL O_2	-1.0	0.52	-1.08	-1.24

[1] $\bar{d}$ - average depth

% Volume E - epilimnion volume expressed as a percentage of the entire lake volume; epilimnion is defined as 0 to 10 m, and hypolimnion as 10 m to bottom (Thienemann, 1928)

O_2 H/E - oxygen coefficient. Hypolimnion oxygen concentration divided by epilimnion oxygen concentration

Sat. O_2 bottom - average oxygen saturation at bottom of lake

ΔE - difference between actual oxygen concentration and the amount that could be present at full saturation, epilimnion

ΔH - difference between actual oxygen concentration and the amount that could be present at full saturation, hypolimnion

ΔH + E - oxygen deficit for entire lake

[2] Data obtained from computer files of surveillance data stored at the Canada Centre for Inland Waters, Burlington, Ontario.

Source: Based on Thienemann (1928).

Table 3. Average Lake Erie Central Basin Hypolimnion Characteristics for 1973 to 1980

Year	Thickness (m)	Temperature (°C)
1973	4.1 ± 1.0	12.0 ± 1.8
1974	5.0 + 1.0	11.5 ± 2.5
1975	7.1 ± 0.6	8.1 ± 1.9
1976	4.8 ± 2.6	11.6 ± 3.0
1977	4.1 ± 2.1	11.1 ± 0.6
1978	5.6 ± 1.2	11.6 ± 1.7
1979	4.2 ± 1.5	14.1 ± 4.6
1980	5.7 ± 0.5	12.8 ± 0.3

Data from Fay and Herdendorf (1981).

Table 4. Dissolved Oxygen Characteristics of the Central Basin of Lake Erie, 1930 to 1980

	Estimated area of anoxic hypolimnion for Central Basin Lake Erie		Net oxygen demand for Central Basin Lake Erie			
Year	Area (km^2)	% of Hypolimnion	Rate/unit/area mg O_2/m^2/day		Rate/unit/volume mg O_2/L/day	
1930	300[1]	3[1]	0.008[2]		0.054[2]	
1940			0.015[2]		0.067[2]	
1950			0.025[2]		0.070[2]	
1959	3 600[1]	33[1]				
1960	1 660[1]	15[1]	0.037[2]		0.093[2]	
1961	3 640[1]	33[1]				
1964	5 870[1]	53[1]				
1967	7 500[1]	68[1]				
1970	6 600[1]	60[1]	0.039[2]	0.043[4]	0.130[2]	
1972	7 970[1]	72.5[1]				
1973	11 270[1]	92.7[1]	0.023[2]	0.053[3]	0.120[2]	0.120[3]
1974	10 250[1]	87[1]	0.047[2]	0.060[3]	0.110[2]	0.130[3]
1975	400[2]	4.1[2]	0.067[2]	0.067[3]	0.120[2]	0.100[3]
1976	7 300[3]	63[3]		0.075[3]		0.130[3]
1977	2 870[4]	24.8[4]		0.058[4]		0.130[4]
1978	3 980[5]	71.9[5]				0.110[5]
1979						
1980	4 330[5]	35.9[5]				

[1] IJC. 1975. Great Lakes Water Quality 1974, Annual Report.
[2] IJC. 1975b. Great Lakes Water Quality 1975, Appendix B.
[3] IJC. 1977. Great Lakes Water Quality 1976, Appendix B.
[4] IJC. 1978a. Great Lakes Water Quality 1977, Annual Report.
[5] Fay and Herdendorf (1981).

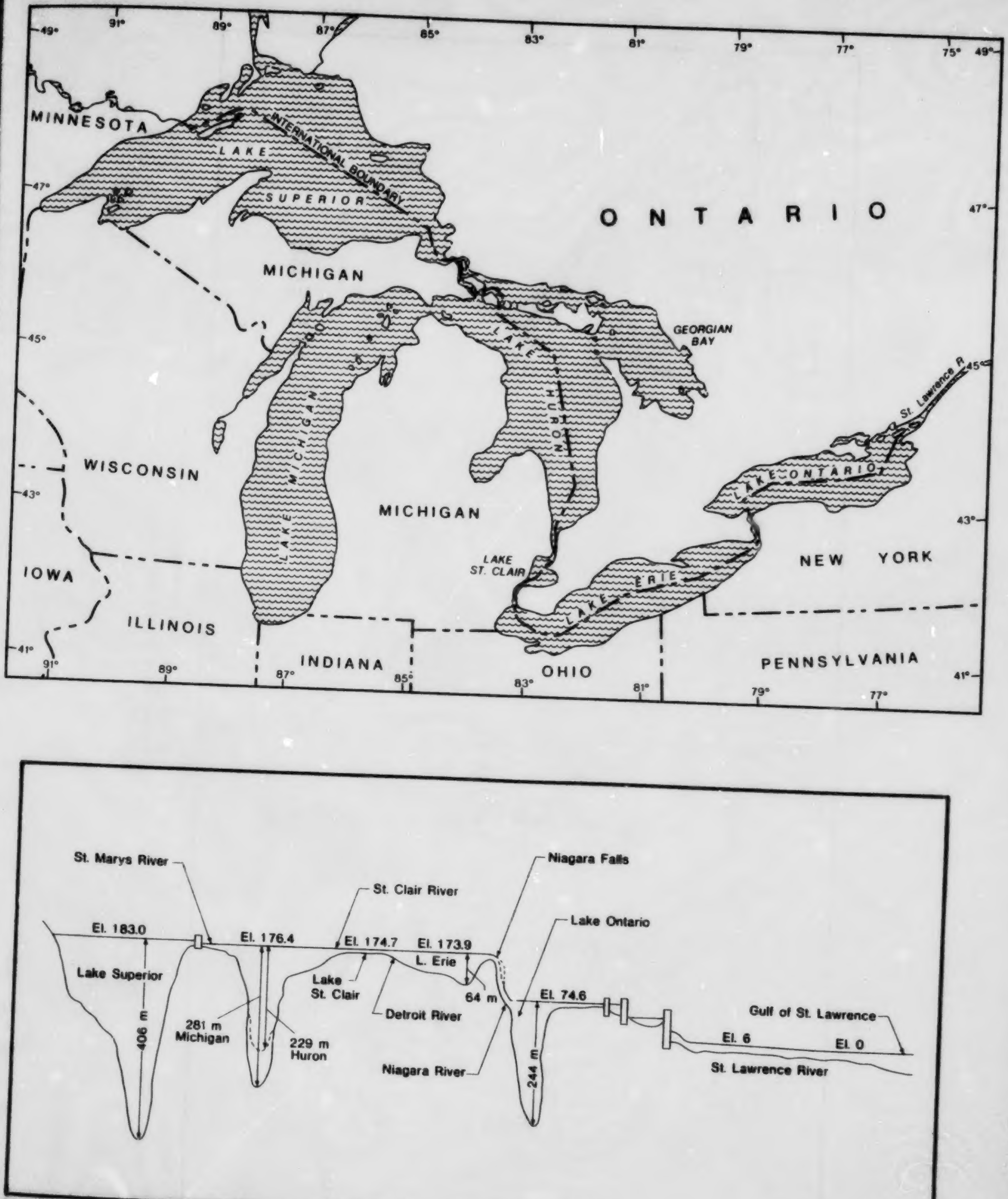

49°
91°
89°
87°
85°
83°
81°
79°
77°
75°
49°
MINNESOTA
INTERNATIONAL BOUNDARY
LAKE
SUPERIOR
47°
ONTARIO
47°
MICHIGAN
GEORGIAN BAY
45°
LAKE
HURON
NORTH
St. Lawrence R.
45°
WISCONSIN
LAKE MICHIGAN
LAKE ONTARIO
43°
MICHIGAN
43°
IOWA
LAKE ST. CLAIR
NEW YORK
LAKE ERIE
ILLINOIS
41°
INDIANA
OHIO
PENNSYLVANIA
41°
91°
89°
87°
85°
83°
81°
79°
77°
St. Marys River
Niagara Falls
St. Clair River
Lake Ontario
El. 183.0
El. 176.4
El. 174.7
El. 173.9
Lake Superior
L. Erie
Lake St. Clair
64 m
El. 74.6
406 m
281 m Michigan
229 m Huron
Detroit River
Gulf of St. Lawrence
El. 6
El. 0
Niagara River
244 m
St. Lawrence River

Figure 1. The Great Lakes.

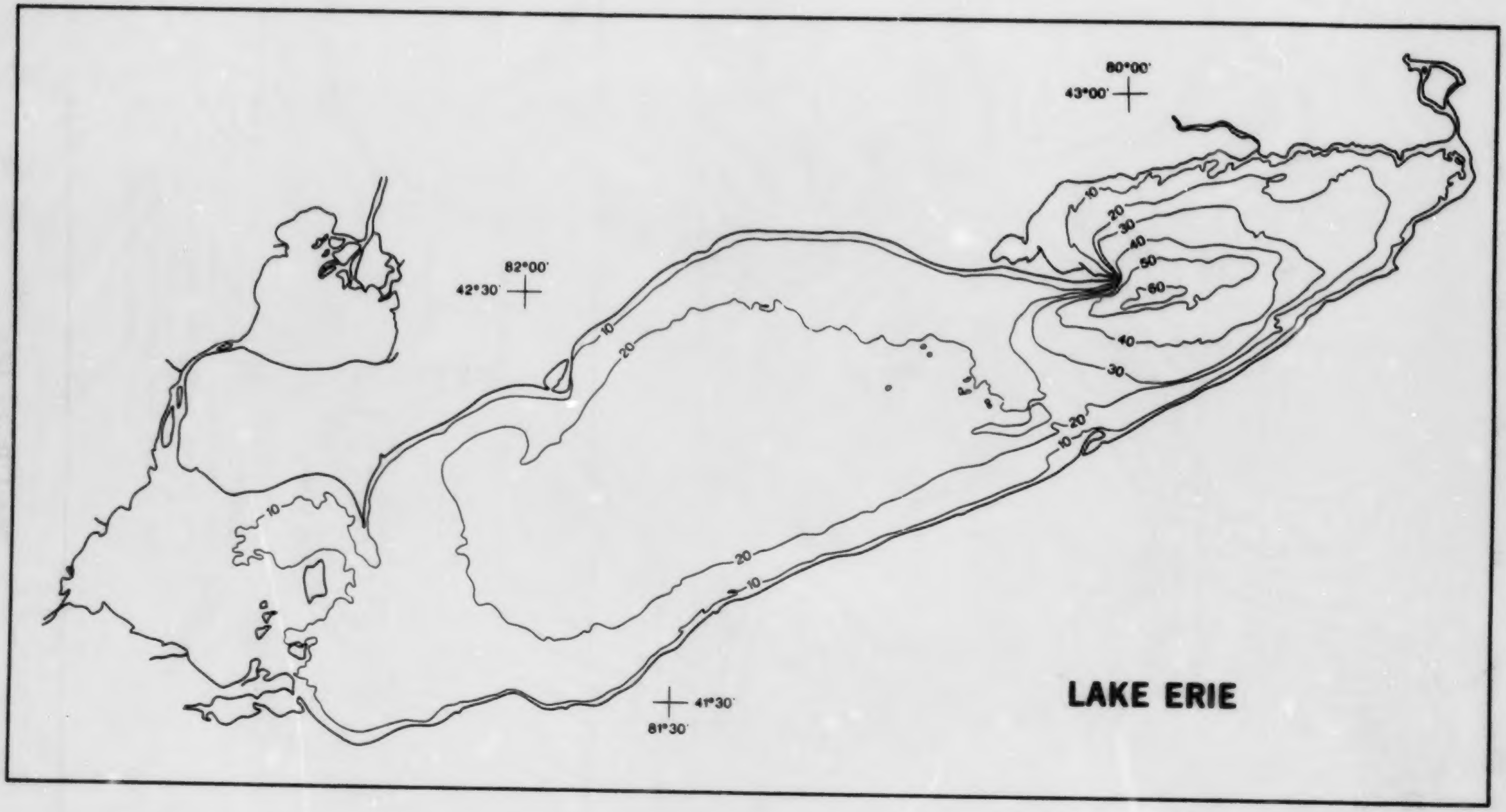

80°00'
43°00'
82°00'
42°30'
41°30'
81°30'
10
20
30
40
50
60
40
30
20
10
10
20
20
10
10
LAKE ERIE

Figure 2. Lake Erie bathymetry (m).

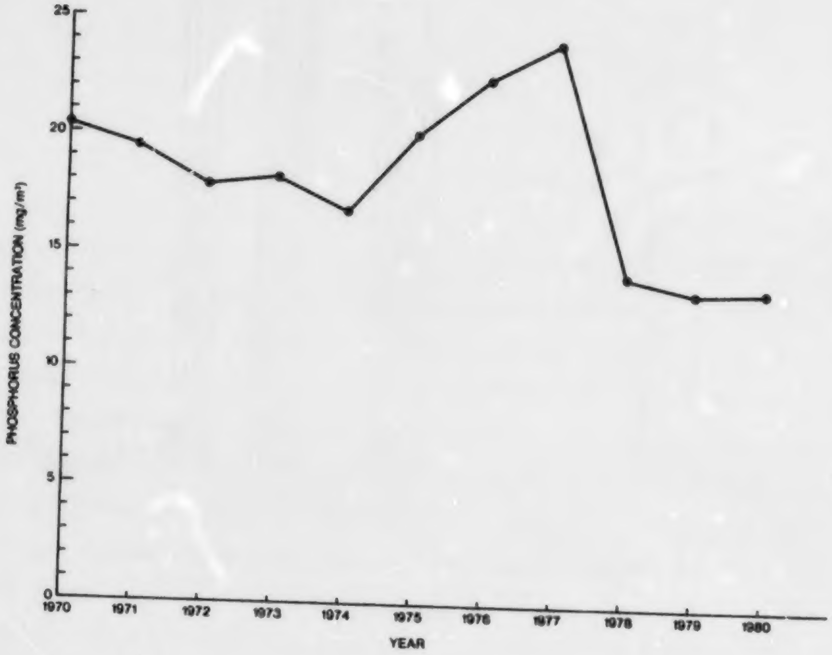


Figure 3. Phosphorus concentration level of Lake Erie, 1970–1980.

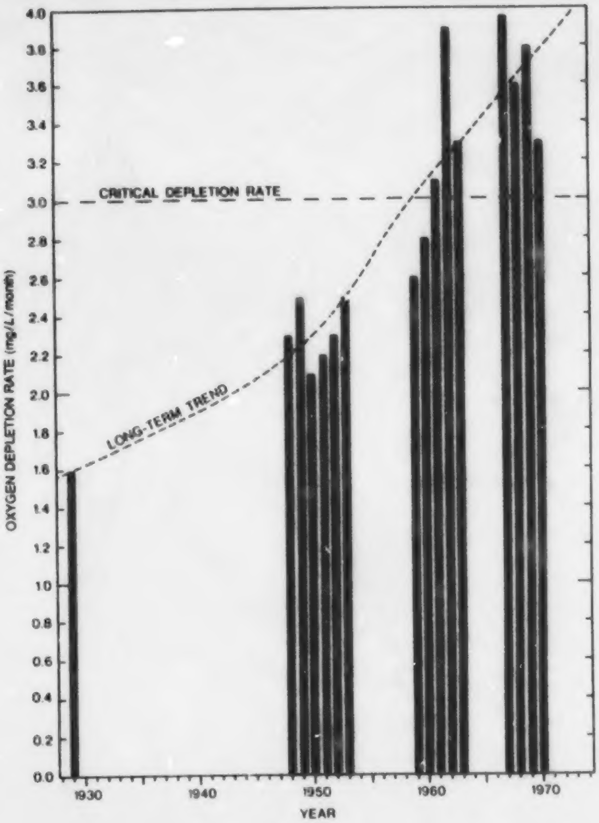


Figure 4. Mean depletion rates for dissolved oxygen during summer in the bottom-water of central Lake Erie. (Courtesy of Dobson and Gilbertson, 1971)

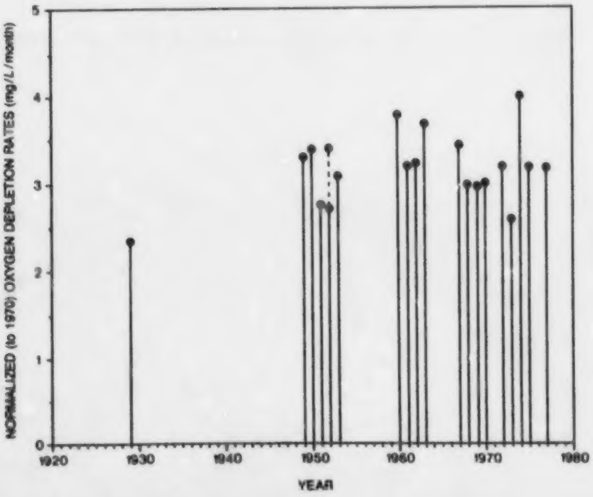


Figure 5. Historic oxygen depletion rates in the hypolimnion of central Lake Erie normalized to thickness conditions of 1970. (Courtesy of Charlton, 1980a)

CHAPTER 3

Temporal Changes in Lake Erie

by A.H. El-Shaarawi

INTRODUCTION

Any attempt to discover the existence and the form of any trend that may be present in the values of a specific water quality indicator for a large body of water like Lake Erie would be hampered by a number of problems. First, large lakes maintain a high degree of spatial variability (El-Shaarawi and Shah, 1978; El-Shaarawi and Kwiatkowski, 1977; El-Shaarawi and Esterby, 1981; El-Shaarawi et al., 1981). The effect of such variability is to decrease the ability of discovering the existence of a trend by adding an additional component to the variance of any statistic that might be employed for trend detection. Moreover, the spatial variability is not constant throughout the year but varies seasonally (El-Shaarawi, 1982). Secondly, most limnological variables possess a typical seasonal cycle, the shape of which varies from year to year. Hence, if the trend is regarded as changes from year to year, then the problem of trend analysis is not simply to follow the yearly changes in a single value such as the mean but to follow the yearly changes in a sequence of curves, each curve representing the seasonal cycle of a year. Thirdly, the design used for collecting the data is problematic. The word "design" is used here to refer to the strategy employed in choosing the locations of the sampling stations and the time for conducting the data collection. Proper trend evaluation requires that the spatial and seasonal effects be estimated from the data, and to obtain adequate estimates of these components an adequate sampling design for collecting the data is necessary. The word "adequate" is defined in terms of the amount of information produced by the design that is relevant to the purposes of the data collection. For example, a sampling design with the location of the sampling stations fixed for all data collection is more appropriate for trend evaluation, while a sampling design which continuously changes the position of the stations is adequate for detecting areas where the water quality objectives are violated. Finally, a sampling design which is a hybrid of the fixed stations and the variable stations designs yields information on the spatial and temporal changes in the lake. Examination of the pattern of the sampling stations used for Lake Erie indicates that a variable and inconsistent sampling strategy was used for collecting the data. This makes the problem of making inferences about the temporal changes in the eutrophic status of the lake difficult. Also it should be mentioned that the sampling of Lake Ontario was performed using the permanent station strategy since 1974, and hence the problem of trend evaluation in Lake Ontario is much simpler than that of Lake Erie.

Chapter 3 presents (a) statistical methods that may be used for discovering the existence and determining the shape of any trend that may be present in the values of a water quality parameter and (b) the applications of these methods for determining the trend in the following parameters: (1) the water level of Lake Erie for the years 1900 to 1979; (2) Niagara River flows for the period 1860 to 1975, and (3) air and water temperature, Secchi disk depth, turbidity, total phosphorus, soluble reactive phosphorus, chlorophyll a, and nitrogen for the period 1967 to 1980. The analysis is performed for the Western, Central and Eastern basins of the lake. The trend in dissolved oxygen depletion rate in the hypolimnion is discussed in Chapters 4 and 5.

DATA SOURCES

Yearly water level and Niagara River flow data were supplied by the Water Planning and Management Branch, Department of the Environment. The water level represents the average of four water level gauges which are located at Buffalo, Cleveland, Toledo and Port Stanley. The Niagara River flow data were measured at Queenston.

The estimates of the monthly Central Basin air temperature were supplied by W.M. Schertzer and are presented also in Lam, Schertzer and Fraser (1983).

The data on the other parameters were obtained from Canada Centre for Inland Waters (CCIW) data files. Corrected chlorophyll a analysis on surface water samples is used in this report. A Whatman GF/C glass fibre filter was used for chlorophyll a extraction and the results were recorded as micrograms per litre.

An electric bathythermograph trace in degrees Celsius was obtained from surface to bottom with a Guideline EBT model 8031 B probe using an HBX-Y recorder. A reversing thermometer was used to verify the EBT temperature.

The data used for the chemical parameters represent the concentration in water at the surface. The technique used for the analysis of water samples is that outlined in the Analytical Methods Manual (Environment Canada, 1975) by chemists at CCIW laboratories and the results are expressed as milligrams per litre.

STATISTICAL METHODS

Reduction of Spatial Variability

Since the sampling strategy used for collecting the data from Lake Erie did not use the fixed station design, it is not possible to follow the temporal changes at a fixed location or at a number of stations in the lake without eliminating the spatial variability. To reduce the effect of the spatial variability on the accurate determination of trend and the shape of the seasonal cycle, the lake can be divided into zones or regions of "homogeneous" waters. This can be done either subjectively on the basis of the general knowledge about the geomorphology of the lake or by using an objective classification procedure such as that developed by El-Shaarawi and Shah (1978). The natural subjective classification is to divide Lake Erie into three geographic regions, the Eastern, Central and Western basins, as shown in Figure 6. The Central Basin is separated by a rocky island chain to the west, and by a low wide sand and gravel ridge to the east. The analysis of trend considered here is restricted to the three geographic zones. To evaluate the efficiency of any classification procedure for reducing the spatial variability, the following method can be used. Let x_1, $x_2 \ldots x_n$ be the values of an observed limnological random variable such as temperature at n sampling stations during a specified cruise. The total variability of the data is given by:

$$TS = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$= \sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2$$

The classification procedure will divide the total data set (stations) into groups (zones), which will result in the division of the lake into k zones. Let x_{ij} be the observed value from the ith station in the jth zone where $(i = 1,2 \ldots n_j)$, n_j is the number of stations in the jth zone $(j = 1,2 \ldots k)$ and $\sum n_j = n$. Similarly, the total variability within the jth zone is

$$TS_j = \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2, \; j = 1,2 \ldots k$$

where $\bar{x}_j$ is the mean of the values in the jth zone. The amount of unexplained variability by the classification is

$$TS. = TS_1 + TS_2 + \ldots + TS_k$$

As a measure of the effectiveness of the classification in reducing the total spatial variation, we use the quantity

$$EC = (1 - TS./TS) \times 100$$

The values of EC lie between 0 and 100, the closer this value to 100, the more effective the classification. In fact, under the assumption that $x_1, x_2 \ldots x_n$ are normally distributed with a common mean and variance, the random variable TS./TS is distributed as the beta distribution with parameters $(n-k)/2$ and $k/2$. Hence, the mean and the standard deviation of EC are

$$E[EC] = (k/n)100$$

and

$$\sigma[EC] = (\sqrt{k(n-k)/(n+1)}\,/n)100$$

respectively. Table 5 presents the values of EC for the geographic zones (Eastern, Central and Western basins), using Lake Erie's surface temperature data from 1968 to 1976. Also, these values are plotted in Figure 7 against months. The plot shows that the value of EC increases with time, starting at March until July, stabilizes at its highest level in August, September and October, and then decreases for the months of November and December. This shows that the zonation is the most useful during midsummer to mid-fall.

Determination of the Seasonal Cycle

Once the lake is divided into zones, the seasonal cycle for each zone can be estimated using regression by regarding the limnological variable as a dependent variable and time in Julian days as an independent variable. The resultant regression equations for the different zones can be tested for equality. If the equality of the regression equations is accepted, then a single seasonal cycle for the entire lake can be obtained. Let x_{ijt} be the value of the variable of interest in the jth region, from the ith station and during tth cruise where $j = 1,2...k$; $i = 1,2...n_{jt}$; $t = 1,2...T$; n_{jt} is the number of stations in the jth region during the tth cruise; and T is the number of cruises conducted during the year under study. Suppose that the seasonal cycle in the jth zone for the limnological variable can be represented by the function $f_j(t)$ which contains p unknown parameters and such that $p \leq k$. The word "parameter" is used here in the statistical sense; it should not be confused with its use in the field of limnology. For example, chlorophyll a is considered a parameter by limnologists but here is considered a variable. The adequacy of $f_j(t)$ to represent the seasonal cycle can be tested by noting that the quantity

$$PES_j = \sum_{t=1}^{T} \sum_{i=1}^{n_{jt}} (x_{ijt} - \bar{x}_{jt})^2$$

represents the "pure" error sum of squares, and the quantity

$$RSS_j = \sum_{t=1}^{T} \sum_{i=1}^{n_{jt}} (x_{ijt} - \hat{f}_j(t))^2$$

represents the residual sum of squares from the fitted model where $\hat{f}_j(t)$ is the estimate of $f_j(t)$. The lack of fit sum of squares is then defined as

$$LF_j = RSS_j - PES_j$$

The suitability of $f_j(t)$ to describe the seasonal pattern can be evaluated using the statistic

$$F_j = (J_j - T)\ LF_j/(T-p)\ PES_j$$

The statistic F_j has an F-distribution with $(T-p)$ and $(J_j - T)$ degrees of freedom, where

$$J_j = n_{j1} + n_{j2} + \ldots + n_{jT}$$

Detection of Trend

The approach given here for discovering the existence of trend in the data is to choose a model for the seasonal cycle and then to estimate the unknown parameters of the model for each individual year. This is followed by testing the hypothesis that these parameters are equal for all of the years. If the test results in the acceptance of this hypothesis, we then conclude that there is no evidence of trend in the data. On the other hand, if the test rejects the hypothesis of equality, the problem is to estimate the pattern of change and to relate it to other factors.

WATER LEVEL

The trace of the yearly mean water level for the Central Basin of Lake Erie for the period 1900 to 1979 is shown in Figure 8. The minimum water level for this period was 568.08 ft, which was reached in 1934, while the maximum water level was 572.51 ft, which occurred in 1974. Hence, the fluctuations in the mean yearly water level exceeded 4 ft during the 80-year period. The graph shows that the water level is nonstationary, that is, the water level series is subject to systematic changes. Two basic features of the nonstationarity are (i) irregular cyclic changes and (ii) a rise in the mean water level after 1944. To smooth the data, the five-year moving median was calculated, as shown in Figure 9. This figure indicates that the water level was subject to cyclic variabilities of high frequency prior to 1934 and thereafter by low frequency cycles with larger amplitude.

A Model for the Water Level

Regression methods are used to develop an adequate empirical model for the water level of Lake Erie. The following steps were followed to build the model: (i) an initial model was fitted to the data; (ii) the adequacy of the model was investigated by examining the residuals and by moving regression; (iii) the model was modified according to the results of (ii); (iv) the modified model was then fitted to the data and steps (ii) and (iii) were repeated; and (v) this process was continued until an adequate model was found.

The basic statistical methods used for developing the model will be used repeatedly in this report; a general account of these techniques is given in the Appendix.

The Model

Since the graphical display indicated the presence of periodicity in the data, the decision was made to start with a model of the form

$$y_t = \alpha_0 + \alpha_1 \sin\omega t + \alpha_2 \cos \omega t + \varepsilon_t \tag{3.1}$$

where y_t is the mean water level of the tth year ($t = 1,2...80$), $t = 1$ corresponds to the year 1900 and $t = 80$, to the year 1979; the parameters α_0, α_1, α_2 and ω are unknown constants; ε_t is a random variable which is assumed to be normally distributed with mean 0 and variance σ^2; and ε_1, $\varepsilon_2 ... \varepsilon_{80}$ are independent.

Since model 3.1 is linear in the parameters α_0, α_1 and α_2, their values can be easily estimated using least squares. However, because ω is also unknown and is nonlinear, the estimation of these parameters can be obtained only by iteration, which is performed in the following manner. Assume first that $\hat{\omega}$ is known and let $\hat{\alpha}_0(\omega)$, $\hat{\alpha}_1(\omega)$, $\hat{\alpha}_2(\omega)$ and $\hat{\sigma}^2(\omega)$ be the least squares estimates for α_0, α_1, α_2 and σ^2, respectively. The maximum likelihood estimate ω for $\hat{\omega}$ can be shown to be the value of ω for which $\hat{\sigma}^2(\omega)$ is minimum. To start the iteration an initial value for ω is needed, and this can be obtained by dividing the number of peaks observed in the water level series by the length of the series and multiplying the result by 2π. The Newton-Raphson method was used to find $\hat{\omega}$. The final estimates of α_0, α_1, α_2 and σ^2 are $\hat{\alpha}_0(\hat{\omega})$, $\hat{\alpha}_1(\hat{\omega})$, $\hat{\alpha}_2(\hat{\omega})$ and $\hat{\sigma}^2(\hat{\omega})$, respectively. Figure 10 gives the plot of the residual sum of squares $77\ \sigma^2(\omega)$ against the period $2\pi/\omega$, which indicates that the cycle has an estimated period of 27.5 yr. The observed water level and the estimated values from model 3.1 are plotted against years in Figure 11. This graph indicates that the initial model does not adequately fit the data. To determine what types of modification in model 3.1 are needed to improve the goodness of fit, the standardized and the orthogonalized residuals (Appendix) are plotted against years in Figure 12. The residual plots show that there is a strong periodicity in the data which is not accounted for by the model. As well, the cumulative orthogonalized residuals are plotted against years in Figure 13. This figure indicates that in the first 30 years water level values were fitted well by the model; thereafter, the cumulative orthogonalized residual decreased steadily up to year 1943, and then increased for the remaining period. This indicates that model 3.1 overestimated the water level for the years 1930 through 1943 and underestimated it for the years 1944 to 1979. To detect inconstancy of variance, the cumulative squared residuals are plotted with their expected values against time in Figure 14. It is clear that the model is not adequate, since all of the cumulative squared residual values fall below their expected values.

A moving regression technique was also used to examine the constancy of the regression parameters with time. The time interval used for the moving regression was set at 20 years and the value of ω was set at $\hat{\omega}$. Figure 15 gives the plot of the estimated values of $\hat{\alpha}_0$ against years, which shows that the $\hat{\alpha}_0$ values are not constant and can be represented by a quadratic equation. Similarly, Figures 16 and 17 present the plots of the estimated values of α_1 and α_2 against years and indicate the inconstancy of these parameters. To take into account the pattern of α_0, model 3.1 is modified to

$$y_t = \alpha_0 + \theta_1(t - \bar{t}) + \theta_2 (t - \bar{t})^2 + \alpha_1 \sin (2\pi t/27.5) + \alpha_2 \cos (2\pi t/27.5) + \epsilon_t \qquad (3.2)$$

where $\bar{t} = 40.5$, the mean of the numbers 1, 2...80; θ_1 and θ_2 are two additional unknown parameters which account for the quadratic pattern of α_0. The effect of adding θ_1 and θ_2 to model 3.1 can be tested using the statistic

$$F = 75 \, (\mathrm{Res}_1 - \mathrm{Res}_2)/2 \, \mathrm{Res}_2$$

where Res_1 and Res_2 are the residual sums of squares under model 3.1 and model 3.2 respectively. The distribution of F is the

F-distribution with 2 and 75 degrees of freedom. The calculated value of F is 17.79, which is highly significant ($p \leqslant 0.1$).

Repeating the previous analysis assuming model 3.1 as the initial model indicated that another periodic component with an approximately 40-year period is needed in the model. Hence, model 3.2 is modified to

$$y_t = \alpha_0 + \theta_1(t - \bar{t}) + \theta_2(t - \bar{t})^2 + \alpha_1 \sin(2\pi t/27.5) + \alpha_2 \cos(2\pi t/27.5) + \beta_1 \sin(2\pi t/40) + \beta_2 \cos(2\pi t/40) + \varepsilon_t \qquad (3.3)$$

where β_1 and β_2 are included in the model to account for the additional periodic components. The inclusion of those two parameters substantially improved the fit, and the associated F-statistic for testing the importance of β_1 and β_2 is

$$F = 73\ (\mathrm{Res}_2 - \mathrm{Res}_3)/2\ \mathrm{Res}_3$$

where Res_3 is the residual sum of squares under model 3.3. The value of F is 5.23, which is significant at the 1% level when compared with the critical value of the F-distribution with 2 and 73 degrees of freedom. Figure 18 presents the observed water level values and their estimated values from model 3.3. A summary of the process used in attaining the final model is given in Table 6.

Seasonal Changes

Figure 19 illustrates the seasonal variabilities of the water level. The figure presents the mean water level for spring, summer, fall and winter. These means are calculated using the data for the period 1965 to 1979 only. The lake has the highest water level during the summer, while the second highest values occur during the spring. The fall water level exceeds those for the winter months. It is evident from the figure that the differences between the summer values and the spring values are larger for the years 1965 to 1973 than those for the years 1974 to 1979.

NIAGARA RIVER FLOWS

Figure 20 gives the yearly mean of the Niagara River flows at Queenston in thousands of cubic feet per second for the period 1860 to 1975. These values represent the yearly mean outflow from Lake Erie. The general pattern of this time series clearly resembles that of the lake water level. This is shown in Figure 21 where the Niagara River flows are plotted against Lake Erie water levels for the period 1900 to 1975. This figure shows that the relationship between the two variables is linear. The regression line is also indicated in the figure and the regression equation is

$$y = 198.132 + 20.677x$$

where y is the flow and x represents the water level. The value of $R^2 = 0.97$ and the estimated standard deviation of y is 3.469 cubic feet per second after accounting for the effect of the water level. Hence the increase in the lake water level by one foot will result in increasing the lake outflow by 20 677 cubic feet per second.

AIR TEMPERATURE

Table 7 shows the Central Basin monthly (April to November) mean air temperature in degrees Celsius for the years 1967 to 1978. The seasonal pattern of temperature is estimated as the average of the monthly means (given in Table 7) and shown in Figure 22a. From this figure, it can be seen that the maximum temperature occurs during August.

To discuss the year to year differences in the monthly (April to November) mean air temperature, the following procedure is used. Let x_{id} be the mean temperature for the month i (for i = 1,2...8) during the year d(for d = 1,2...12). Let $\overline{x}_i$ be the average of the means for the ith month, i.e.,

$$\overline{x}_i = \frac{x_{i1} + x_{i2} + \ldots + x_{i12}}{12}$$

Define the residual difference between the ith mean in the dth year and $\overline{x}_i$ as

$$R_{id} = x_{id} - \overline{x}_i$$

The pattern of R_{id} indicates the nature of a change in a particular year. For example, in Table 8, since the year 1973 shows positive residuals for all months except for the month of November, it can be regarded as warmer than average. As the year 1972 shows negative residuals except for the month of May, it can be considered colder than usual.

Another method for displaying the differences between years is to form the CUSUM (cumulative sum) graph. The CUSUM is defined as follows:

$$C_{id} = C_{(i-1)d} + R_{id}$$

(for i = 1,2...8), where C_{id} is the CUSUM for the dth year at the ith month and $C_{0d} \equiv 0$. The graphs of the CUSUM for the years 1967 to 1978 are shown in Figure 22b and the values of C_{id} are given in Table 8. For a normal year, the graph of the CUSUM should be very close to the x-axes, while for a warmer year, the graph should show a positive drift and for a colder than average year, a negative drift. The results are summarized below and in Table 9 for each year:

1967 - This year started as normal for April, May and June, but showed colder summer and fall. As can be seen from Table 8, the CUSUM plot is close to the x-axes for April, May and June, but shows a drift in the negative direction, indicating that the summer and fall were colder than average.

1968 - The residuals in Table 8 show that this year has a normal spring, a colder summer, but a warmer fall. However, the CUSUM plot shows that these changes were not very severe.

1969 - This can be regarded as a normal year. The residual values are not very high. The period July to October showed above normal temperature. The CUSUM plot does not show strong evidence of deviations from the norm.

1970 - Clearly, this year is warmer than average, since all residual values were positive except that of the month of July. The CUSUM graph drifts in the positive direction, showing an above average temperature in this year.

1971 - Residual values show below average spring and summer temperature and above average fall temperature. The CUSUM plot does not indicate strong deviations from the norm.

1972 - Residual values show that this year was colder than usual. In fact, there was only one case with positive residuals and this occurred in May. The CUSUM plot for this year shows the same result. It drifts in the negative side of the x-axes.

1973 - Residual values and the CUSUM pattern for this year show that this year was warmer than usual.

1974 - All residual values are negative except that of April. Hence, this year has colder temperatures than average, as indicated by the CUSUM plot.

1975 - Residuals in Table 8 and the CUSUM plot indicate that the summer temperature values were above normal.

1976 - This year shows a colder summer and fall.

1977 - This year temperature appears to be very close to the norm.

1978 - Residual values indicate that this year was warmer than average.

WATER TEMPERATURE

The Seasonal Cycle of the Surface Water Temperature

The seasonal cycles of the surface water temperature for the Western, Central and Eastern basins of Lake Erie are shown in Figure 23. Figure 23 is based on the mean surface temperature data for the cruises conducted during 1967 and 1968. The graph shows that the maximum temperature occurred between day 210 and day 240 in the month of August. Also, the temperature for the Western Basin exceeds that of the Central and Eastern basins until day 260, after which this pattern is reversed.

The year can be divided arbitrarily into three periods: (1) the warming period; (2) the stagnation period; and (3) the cooling period. The warming period is taken between day 1 and day 200 and the cooling period, between day 250 and day 365. The summer stagnation period occurs during July and August (day 200 to day 250) when the lake has approximately uniform surface temperature. Figure 24 presents the plot of the mean surface temperature for the Western Basin (y-axes) against the mean surface temperature for the Central and the Eastern basins, in which the 45° line represents equal temperature values. During the warming period, all the points fall above the 45° line. This indicates that the Western Basin is warmer than the other two basins, with the Eastern Basin having the coldest temperature. However, the differences in temperature between the three basins decrease with increasing time (and thus temperature). During the cooling period, the Western Basin is colder than the other two, and the difference between the three basins increases with decreasing temperature and hence increasing time. Also, it can be seen from the graph that during the warming period the maximum difference between the mean temperature of the Western Basin and the Central Basin was about 8 degrees; this occurred when the Western Basin temperature was about 12°C. In the cooling period, the maximum temperature difference was about 4°C; this occurred when the Western Basin surface temperature was about 4°C.

Statistical Analysis of the Surface Temperature Data

The purpose of the following analysis is to develop statistical procedures for estimating the yearly seasonal temperature cycle and for determining the changes in its form from year to year. Let y_{ts} be the mean of the observed surface temperature for the tth year ($t = 1,2...m$) and on the sth Julian date ($s = 1,2...365$), where m is the number of years under study. It is assumed that the seasonal cycle of the temperature in the tth year can be presented by the model

$$y_{ts} = \alpha_t + \beta_t \cos ws + \gamma_t \sin ws + \epsilon_{ts} \qquad (3.4)$$

where $w = 2\pi/365$; α_t, β_t, γ_t are unknown parameters; and ϵ_{ts} is a normal random variable with mean 0 and variance $\sigma_t{}^2/n_{ts}$, where $\sigma_t{}^2$ is the variance of a single temperature measurement and n_{ts} are the number of temperature measurements used in calculating the mean y_{ts}.

This model was selected because the observed seasonal cycle for temperature (Fig. 25) takes the shape of a cosine function. Another way of expressing model 3.4 is

$$y_{ts} = \alpha_t + R_t \cos \left(\frac{2\pi s}{365} + \phi_t\right) + \epsilon_{ts} \qquad (3.5)$$

where R_t is the amplitude and ϕ_t is the phase, and these are given in terms of β_t and γ_t by the relation

$$R_t = \sqrt{\beta_t^2 + \gamma_t^2}$$

and

$$\tan \phi_t = \gamma_t / \beta_t$$

The quantity $\alpha_t + R_t$ represents the estimated maximum temperature in the tth year, while the quantity $D_t = -365\ \phi_t/2\pi$ represents the time at which the maximum temperature has occurred. Model 3.4 is preferable to model 3.5 for the statistical analysis, since it is linear in the unknown parameters α_t, β_t and γ_t, and hence these parameters can be estimated by the direct application of the theory of least squares. However, the parameters R_t and ϕ_t of model 3.5 have a more meaningful interpretation. Table 10 summarizes the results of fitting model 3.4, and hence model 3.5, to the mean surface temperature values for each of the Western, Central and Eastern basins of Lake Erie for the years 1967 to 1978. The results shown for each basin are the estimate of the maximum temperature for each year, the estimated phase ϕ_t, the estimated number of Julian days D_t for reaching the maximum temperature, the coefficient of determination ${R_t}^2$ and the F-statistic for testing the significance of the regression. In the tables all the values for the F-statistic are marked by two asterisks (**) to indicate that the regression equations are significant at the 1% level. Also, the smallest value obtained for the coefficient of determination R^2 is 0.76. The maximum estimated surface water temperature in the Western Basin ranged from 21.47°C in 1971 to 25.60°C in 1973. The corresponding values for the

Central Basin were 19.46°C in 1971 to 23.90°C in 1973. This agrees with the results noted earlier about the mean air temperature for 1973 and 1975 in the Central Basin. The Eastern Basin estimated minimum (19.78°C) of the maximum temperature was obtained in 1969, while the maximum of the estimated maximum temperatures (23.78°C) occurred again in 1973.

The values of D_t in Table 10 show also a large degree of variation. In the Western Basin, the smallest number of days D_t for reaching the maximum temperature was 203.2 Julian days, and this occurred in 1971. While the maximum of D_t was 221.4, which occurred in 1969, the corresponding values for the minimum and maximum of D_t for the Central Basin were 206.8 in 1967 and 229 in 1969, respectively. On the other hand, the corresponding values for D_t for the Eastern Basin were 209.2 in 1971 and 232.8 in 1968. The median Julian days D_t were 217.2, 226.1 and 226.6 for the Western, Central and Eastern basins, respectively. From this it can be concluded that the Western Basin reaches its maximum nine days earlier than the Central Basin. The difference between the median of D_t of the Central and Eastern basins is very small.

Year to Year Variation in the Surface Water Temperature

Model 3.4 can be used to determine whether the year to year variabilities in the surface temperature seasonal cycle exceed those expected on the basis of statistical fluctuations. This is equivalent to testing the hypothesis

$$H: \quad \alpha_t = \alpha, \ \beta_t = \beta \text{ and } \gamma_t = \gamma$$

for all values of t. This means that the parameters of model 3.4 are independent of years, and hence one single equation can be used to represent the temperature adequately. This can easily be done using the Fisher F-statistic for testing the equality of several regression equations. Due to the inequality of the values of $\hat{\alpha}_t^2$, the analysis was performed using the empirical weights $\hat{\alpha}_t^2$, which is the estimated value of α_t^2. Table 11 summarizes the results of this analysis, which indicates that all the F-values were significant at the 1% level for each basin.

To determine the pattern and overall magnitude of variability from year to year, Figures 26a, 26b and 26c present the plot of the yearly mean residuals against years for the Western, Central and Eastern basins, respectively. Residuals are given as the difference between the observed value and its estimate from model 3.4 and the yearly mean residual is then the average of the residuals for each year. A positive mean residual is indicative of a warmer year and a negative value is indicative of a colder year. These graphs indicate that 1972 and 1975 were the coldest years, while 1971 and 1973 were the warmest.

Vertical Variation in Temperature

During the summer Lake Erie can be classified vertically on the basis of temperature and hence water density, into three distinct layers: the epilimnion, the thermocline or the mesolimnion, and the hypolimnion. Figure 27 shows the temperature depth profiles for the cruise of July 29 to August 3, 1968. The epilimnion comprises the least dense upper layer of water where a gradual temperature change occurs. Below this zone lies the thermocline where a rapid temperature and density gradient exists. Finally, the deepest zone, the hypolimnion, contains the most dense water at a relatively constant temperature. The average depth of the hypolimnion in the Central Basin is about 4 m, although it does vary both spatially and temporally. In the Eastern Basin, the depth of the hypolimnion is much greater due to the greater depth of the Eastern Basin, and in the Western Basin, no hypolimnion exists. During the winter season, the reverse is true. For the colder water mass to move from the bottom layer of the lake in the summer to the top layer in the winter, at some point between mid- to late fall the entire lake must become isothermal. This phenomenon is referred to as the fall overturn. Similarly, the winter stagnation period is followed by the spring overturn.

Figures 28 and 29 give the temperature depth profiles for each station sampled in the Central Basin in 1967 and 1978, respectively. Two observations can be made here: (i) the epilimnion, hypolimnion and thermocline are not as well defined and (ii) there is a difference between the boundaries of the zones for 1967 and 1978. The difference between the boundaries of the layers is due to the difference between the pattern of temperature for the two years. The year 1967 is colder than average, while 1978 is warmer than average, as shown previously. Such differences are important in discussing changes in water quality.

Year to Year Variability in the Hypolimnion Temperature (Central Basin)

Figure 30 presents the observed mean temperature values for the epilimnion and the hypolimnion zones in the Central Basin. The stratification period for the Central Basin is taken as between 150 and 250 Julian days. There is a very good separation between the epilimnion and the hypolimnion. The minimum mean temperature for the hypolimnion was about 6°C, while the maximum mean value was slightly above 13°C. Also, it was noted that the hypolimnion temperature rises slightly with time. Temperature values were fitted to the linear regression model

$$y_t = \alpha + \beta t + \varepsilon_t$$

where y_t is the mean temperature for the Central Basin at the tth Julian day; α, β are the slope and the intercept of the line, respectively; and ε_t is a normal random variable with mean 0 and variance σ^2. The fitted regression equations are shown in Figure 30, and the estimated values for α and β are given for each year in Table 12. The figure shows that the coldest hypolimnion was obtained in 1975 and the second coldest, in 1978. The warmest hypolimnion was found in 1977. Furthermore, the estimate of the slope, β, gives some indication of the degree of the transport of heat to the hypolimnion from the other layers. The values of β are very close to zero for 1967, 1968 and 1973, and hence the transport of heat was minimum in these years, whereas in 1969, 1971, 1978 and 1979 there was a modest degree of heat transport ($.025 < \beta < .05$) and the remaining years experienced a higher degree of heat transport ($.05 < \beta < .09$).

CHLOROPHYLL a

Figure 31 shows plots of the mean surface uncorrected chlorophyll a against time in Julian days for the Western, Central and Eastern basins of Lake Erie. The data are available for 1968, 1970 and 1972 for the Western Basin; for 1968, 1970, 1979 and 1980 for the Central Basin; and for 1968, 1970, 1972 and 1980 for the Eastern Basin. This figure shows that the Western Basin has the highest phytoplankton biomass whereas the Eastern Basin has the lowest. The difference between the biomass of the Western and the Central basins is much more pronounced than that of the Eastern and the Central basins. The year 1970 has the highest chlorophyll a peaks, while the 1980 seasonal cycle falls below those of the other years.

Western Basin

The results of fitting model 3.4 to the Western Basin data are given in Table 13. They show that the coefficient of determination is quite high (0.74). An increasing trend in the values of the estimated maximum is apparent. To test whether chlorophyll a has the same seasonal cycle but with different initial conditions, model 3.4 is specialized to

$$y_{ts} = \alpha_t + \beta \cos wt + \gamma \sin wt + \varepsilon_{ts} \qquad (3.6)$$

The hypothesis tested is then

$$H_1: \ \beta_t = \beta \text{ and } \gamma_t = \gamma \text{ for all } t$$

If H_1 is accepted, that is, model 3.6 provides a reasonable fit to the data, then the next hypothesis tested is

$$H_2: \ \alpha_t = \alpha \text{ for all } t$$

If H_2 is also accepted, then it can be concluded that there are no differences between years. The statistics F_1 and F_2 are used to test H_1 and H_2, respectively, where

$$F_1 = (n - 3m)\ (RES_1 - RES)/2(m-1)RES$$

and

$$F_2 = (n-m-2)\ (RES_2 - RES_1)/(m-1)RES_1 \tag{3.7}$$

The quantities n and m denote the total number of cruises and the number of years, respectively (while RES, RES_1 and RES_2 represent the residual sum of squares under model 3.6 for H_1 and H_2, respectively). Under the assumptions of these models, F_1 and F_2 have F-distribution with 2(m-1), (n-3m) and (m-1) and (n-m-2) degrees of freedom, respectively. Furthermore, it is possible to test the hypothesis that all years have the same seasonal cycle. The hypothesis tested is

$$H_0: \quad \alpha_t = \alpha,\ \beta_t = \beta \text{ and } \gamma_t = \gamma \text{ for all } t$$

The statistic F_0 is used to test H_0 and is given by

$$F_0 = (n-3m)\ (RES_0 - RES)/3(m-1)RES$$

where RES_0 is the residual sum of squares under H_0.

Table 14 gives the values of the statistics F_1, F_2 and F_0 and their associated degrees of freedom. Although none of these tests indicates statistically significant differences from year to year, the values of $\hat{\alpha}_t$ obtained from model 3.6 indicate a systematic increase in the level of chlorophyll a in the Western Basin, as shown in Table 15.

Central Basin

The previous analysis has been repeated for the Central Basin and the results are given in Tables 16, 17 and 18. Table 17 shows that the hypothesis H_1 was not significant. However, the fact that H_2 was significant at the 5% level indicates that the differences between years are significant; hence model 3.6 can be used to represent chlorophyll a in the Central Basin. The values of $\hat{\alpha}_t$ indicate that chlorophyll a was the highest in 1970 and the lowest in 1980. Also it appears that a linear decreasing trend in the chlorophyll a values has occurred since 1970.

Eastern Basin

Tables 19, 20 and 21 give the results of the statistical analysis for the Eastern Basin. Model 3.6 is found to be appropriate for chlorophyll a in the Eastern Basin. This is supported by noting that the value of F_1 is not significant at the 5% level but F_2 is significant at the 1% level, which indicates that β_t and γ_t are the same for all the years while the variation in α_t is statistically significant. The values of $\hat{\alpha}_t$ are given in Table 21 and show a decreasing trend in chlorophyll a values.

In summary, there appears to be a significant decreasing trend in chlorophyll a for the Central and Eastern basins between 1968 and 1980, and an insignificant increasing trend in the Western Basin between 1968 and 1972. These results are shown in Figure 32 where the year effects, $\hat{\alpha}_t$, are plotted against years for each of the three basins.

TOTAL PHOSPHORUS - TP

Figure 33 presents the plot of the mean total phosphorus, TP, against time in Julian days for the surface water of each basin. The general shape of the seasonal cycle for TP is that high values occur in the winter and spring, followed by a rapid decrease in early summer, reaching a minimum late in summer. Also TP values are higher in the Western Basin than in the Central and Eastern basins. The degree of variability from year to year appears to differ from basin to basin. The figure shows that the highest variability occurred in the Western Basin, followed by that of the Central Basin. Furthermore, the shape of the seasonal cycle can be either represented by model 3.4 or by a quadratic function. The same analysis as that used for chlorophyll a was performed.

Western Basin

Tables 22, 23 and 24 present the results of the statistical analysis of TP using model 3.4. Table 23 shows that model 3.6 gives a reasonable representation for the data. The statistic F_2 is highly significant, which indicates that the year to year variability is significant and is determined by the pattern of the variability of α_t. As shown in Table 24 the values of TP for 1977 and 1978 are much smaller than those of 1968 to 1971. In fact, it can be concluded that there was a slight increase in TP when 1968 is compared with 1970 and 1971. Also, it can be noted that a sharp decline in TP occurred in 1972.

Central Basin

The results are summarized in Tables 25, 26 and 27, which show that model 3.6 is appropriate. The values of $\hat{\alpha}_t$ indicate a strong and significant decreasing trend for TP.

Eastern Basin

The TP values for the Eastern Basin indicate a significant decreasing trend. However, the changes are much less pronounced than those found for the Western and the Central basins. The results given in Tables 28, 29 and 30 indicate that the changes are not statistically significant at the 1% level. The values of $\hat{\alpha}_t$ in Table 30 show a decrease after 1971.

Figure 34 presents the plots of $\hat{\alpha}_t$ against years for each of the three basins. This shows that TP values have decreased substantially after 1971, with the Western Basin showing the largest decrease, followed by the Central Basin and the Eastern Basin, respectively. In fact, the concentration of TP in the Western Basin is becoming very close to that of the Central and Eastern basins in 1977, 1978, 1979, and 1980. Figure 35 presents the plot of the values of $\hat{\alpha}_t$ for chlorophyll a against those for TP. The plot indicates a strong positive association between the two, i.e., a decrease in TP is associated with a decrease in chlorophyll a.

SOLUBLE REACTIVE PHOSPHORUS - SRP

Since the discussion of the results is very similar to that given previously, the results for the three basins are combined. The values of F_2 and F_0 (Table 31) are significant at the 1% level for the

Western and the Central basins, and at the 5% level for the Eastern Basin. Since F_1 is not significant for the Western and Eastern basins and significant only at the 5% level for the Central Basin, it can be concluded that model 3.6 is suitable for describing the SRP values. The estimates of α_t are shown in Table 32, which indicate that the concentration of SRP has dropped substantially after 1977 when compared with the SRP concentration prior to 1973. Figure 36 shows the plot of $\hat{\alpha}_t$ against years for each of the three basins.

FILTERED NITRATE NITRITE (NO_3NO_2-N)

Figure 37 shows the plot of the concentration of NO_3NO_2-N against Julian days for the Western Basin and for each year where data are available. Since the NO_3NO_2-N values reach a minimum between Julian days 220 and 280, the minimum occur in August. Model 3.4 is fitted to the data, and the estimated values of the level of NO_3NO_2-N and the amplitude are given in Table 33 as $\hat{\alpha}_t$ and R_t, respectively, where $R_t = \sqrt{\beta_t^2 + \gamma_t^2}$. These are shown for each basin and for each year where the data are available. The values of $\hat{\alpha}_t$ and R_t for the Western Basin exceed those of the Central and Eastern basins, which indicates that the level and the amplitude of the concentration of NO_3NO_2-N are the highest in the Western Basin. Also, the values of $\hat{\alpha}_t$ are the lowest in 1971 and the highest for 1978 and 1980. The rate of increase in NO_3NO_2-N is the lowest in the Western Basin. Table 34 presents the values of F_2 and F_0 for each of the three basins, which indicates that all the values are significant at the 1% level; hence H_0 is rejected. Since F_2 is significant, model 3.6 is not appropriate and thus model 3.4 is the appropriate model to represent the changes from year to year. Figure 37 presents the observed and the estimated values of NO_3NO_2-N from model 3.4 for the Western Basin. From these graphs it appears that the seasonal cycles for NO_3NO_2-N for the years 1967 through 1970 are the same and the 1978 seasonal cycle is similar to that of 1980. This hypothesis is then tested using the appropriate F-statistic and is accepted at the 5% level. The summary of the parameters of the modified model is given in Table 35. This indicates that the years 1978 and 1980 have higher values for NO_3NO_2-N when compared with the other years and 1971 has the lowest level of NO_3NO_2-N. Figures 38a, 38b and 38c give the observed and the estimated values of NO_3NO_2-N for the Western Basin: Figure 38a shows the results for the years 1967, 1968 and 1970; Figure 38b gives the results for 1971; and Figure 38c gives the results for 1978. Similar plots are given for the Central Basin in Figures 39a, 39b and 39c. Figure 40 shows the plot of the observed and the estimated values (from model 3.4) of NO_3NO_2-N for the Eastern Basin.

AMMONIA (NH_3)

The previous analysis was performed on the values of NH_3. The results indicated that model 3.6 is appropriate for the Western and Eastern basins; the data of Central Basin, however, indicated a slight change in the amplitude, which implies that model 3.4 is more suitable. Table 36 presents the values of F_1, F_2 and F_0 for testing H_1, H_2 and H_0, respectively. The results indicate that H_2 is highly significant except for the Western Basin; this means that there are differences from year to year. Figures 41, 42 and 43 give the observed and the estimated values of NH_3 for the Western, Central and the Eastern basins, respectively. These figures show that model 3.6 fits the data well and the decreasing trend in values of NH_3.

Table 37 gives the estimates of the parameter of model 3.6. The values of $\hat{\alpha}_t$ for the Central Basin are larger than their corresponding values for the Western and Central basins. Also, these values decreased substantially after 1977 when compared with those for 1968, 1970 and 1971.

SECCHI DISK DEPTH

A preliminary examination of the relationship between the mean and the standard deviation of Secchi disk values indicated that the data should be transformed to logs prior to the analysis. Table 38 presents the results of fitting model 3.4 to the logs of the data. The Secchi disk data increase with Julian date until the month of July when a maximum is reached.

The columns headed by amplitude in Table 38 give the estimated maximum log Secchi disk value and the column D_t gives the estimate of the number of Julian days to reach the maximum. As expected, the Western Basin has the lowest amplitude followed by the Central Basin and then the Eastern Basin. Also, the Western Basin has the least defined seasonal cycle. This can be shown from the values of the coefficient of determination R^2 or the values of the F-statistic (which measures the significance of the regression). The R^2 are the lowest for the Western Basin and most of the F-values are not significant. This is in contrast with the corresponding values for the other two basins. Figure 44 shows the plot of the estimated log maximum value of Secchi disk for each basin. From the figure it can be seen that the Western Basin is well separated from the other two, and it appears that there is evidence of an increase in the Secchi disk values in 1976, 1977 and 1978.

The same kind of variability is found for the estimate in the number of days D_t to reach the maximum. With the exception of 1977, the value varied from 188 Julian days to 240 Julian days with the maximum reached in the Central Basin prior to that of the Eastern Basin.

Year to Year Variability in Secchi Disk Values

Table 39 gives the F-statistic for testing the equality of the yearly regression equation. The F-statistic is significant at the 1% level for the Central and Eastern basins, but not significant for the Western Basin. In the previous section, it was found that there appears to be an increase in the estimated maximum Secchi disk values, which gives one type of change from year to year.

Another type of yearly variability can be determined from the plot of the mean residuals which is obtained after fitting a single equation to all the data. Figure 45 gives the plot of the data and the estimated values from model 3.4 against time. Figure 46 gives the plot of the mean residuals. It is clear that Figure 46 does not indicate any important changes except perhaps the drop in 1973 values.

TURBIDITY

Turbidity values must also be transformed to logs prior to fitting model 3.4. The results presented are for the years 1967 through 1972, which are the years where data on turbidity are available.

Discussion of the Results of Fitting Model 3.4 to the Turbidity Data

The results of fitting model 3.4 to the log turbidity data can be summarized (Table 40) as follows. The model explains very well the pattern of the seasonal cycle. The values of the F-statistic are highly significant and the corresponding R^2 values are high. The exception is the 1972 data for the Central and Eastern basins of the lake. The turbidity values attain their minimum at about the end of July and early August, and the minimum is reached for the Western Basin before that of the Central and Eastern basins. The general pattern indicates that turbidity in each basin increased between 1969 and 1972. This is much more pronounced when compared with the previous results obtained for Secchi disk depth.

Year to Year Variability in Turbidity Values

Table 41 gives the values of the F-statistic for the differences between years. All these values are highly significant (significant at the 1% level), which indicates that there are differences in the seasonal cycles from year to year.

PHOSPHORUS AND CHLOROPHYLL a ASSOCIATION

It is well known that eutrophication depends on excessive inputs of phosphorus and nitrogen to the lakes (Vollenweider et al., 1980). In Lake Erie, the point source phosphorus load has been reduced from approximately 10 000 metric tons/yr in 1972-73 to 5700 metric tons/year in 1977 (Slater and Bangay, 1980). It is important to determine whether the reduction in the phosphorus loading is accompanied by a corresponding reduction in phosphorus concentration and in the phytoplankton biomass. In the previous section, it has been demonstrated that the concentration of phosphorus and the values of chlorophyll a are decreasing. To relate the changes in chlorophyll a to those of phosphorus, it is important to correlate quantities which are free from both the spatial and the seasonal variabilities. This can be done by using the year effects (the level of the year value after the elimination of the seasonal and spatial effects) from model

3.4 for both variables. The uncorrected chlorophyll *a* year effect values are plotted against those of TP in Figure 47. It appears from the graph that the relationship can be reasonably described by a straight line. The fitted regression equation is

$$y = 1.6 + 0.206174X$$

where X is the concentration of phosphorus in milligrams per litre and y is the phytoplankton biomass in milligrams per litre as measured by chlorophyll *a*.

Also shown on the graph are the average phosphorus year effects for 1971 and 1980 and the corresponding estimated value of the phytoplankton biomass. This indicates that Lake Erie has responded to the phosphorus reduction program. It should also be noted that the filtered nitrate nitrite has shown an increase during the same period.

RATIO OF AMMONIA TO NITRITE AND NITRATE

It was stated in Rockwell *et al.* (1980) that "a high ratio of ammonia to nitrite indicates a recent source of pollution while a low ammonia to nitrate ratio indicates an earlier (i.e., older) input that has subsequently been oxidized." In Table 42, the ratio of ammonia to nitrite and nitrate is given for the years where the data are available and for each of the three basins. In the table the ratio in the Central Basin is higher than that of the Western and Eastern basins and for the years 1967, 1970 and 1971, the ratio is constant within each basin. The Central Basin data show a very strong reduction in the ratio from 0.229 in 1967 to 0.081 in 1980. This is another sign which shows the response of Lake Erie to the phosphorus control program.

Table 5. The Values of EC for the Geographic Zones of Lake Erie

Cruise date	Percent of explained variation
May 14 to May 23, 1968	56.2
June 17 to June 19, 1968	81.5
July 31 to Aug. 3, 1968	75.6
Sept. 1 to Sept. 2, 1968	63.8
Sept. 30 to Oct. 3, 1968	83.8
Nov. 7 to Nov. 10, 1968	34.6
May 31 to June 4, 1969	42.9
July 3 to July 7, 1969	78.0
July 29 to Aug. 1, 1969	75.1
Aug. 26 to Aug. 30, 1969	87.6
Sept. 15 to Sept. 18, 1969	95.9
Oct. 16 to Oct. 20, 1969	94.8
Dec. 8 to Dec. 12, 1969	47.3
April 8 to April 11, 1970	50.6
May 7 to May 10, 1970	41.5
June 3 to June 6, 1970	62.1
July 4 to July 7, 1970	67.1
July 28 to Aug. 2, 1970	97.2
Aug. 25 to Aug. 30, 1970	70.5
Sept. 24 to Sept. 27, 1970	88.2
Oct. 22 to Oct. 26, 1970	52.5
Nov. 26 to Nov. 30, 1970	35.8
Dec. 14 to Dec. 18, 1970	53.4
July 7 to July 8, 1971	53.4
Feb. 7 to Feb. 12, 1971	94.0
March 3 to March 6, 1971	29.8
April 15 to April 18, 1971	52.4
Aug. 18 to Aug. 21, 1971	54.53
Nov. 24 to Nov. 27, 1971	27.52
April 26 to April 28, 1972	38.3
June 7 to June 10, 1972	31.2
June 28 to June 30, 1972	64.2
Aug. 3 to Aug. 5, 1972	95.2
Aug. 29 to Sept. 1, 1972	82.7
Sept. 29 t Oct. 1, 1972	44.3
Nov. 11 to Nov. 14, 1972	54
April 13 to April 16, 1973	73.3
July 26 to July 30, 1973	77.1
Aug. 29 to Aug. 31, 1973	48.5
Nov. 10 to Nov. 13, 1973	72.8
April 24 to April 27, 1974	52.63
Aug. 23 to Aug. 25, 1974	46.34
April 6 to April 10, 1975	63.9
May 13 to May 22, 1975	79.9
June 25 to June 29, 1975	74.9
Aug. 6 to Aug. 11, 1975	69.0
Oct. 8 to Oct. 11, 1975	63.9
Oct. 28 to Oct. 31, 1975	50.8
Nov. 26 to Nov. 30, 1975	45.0
April 10 to April 13, 1976	39.0
May 28 to May 30, 1976	51.3

Table 6. A Summary of the Steps Used in Modelling the Water Level

Model	Equation	S.D.	F
Initial	$Y_t = 570.405-0.661 \sin(\pi t/13.75) + 0.159 \cos(\pi t/13.75)$	0.807	—
Modified	$Y_t = 569.985+0.011(t-40.5) + 0.000787(t-40.5)^2$ $- 0.568 \sin(\pi t/13.75) + 0.167 \cos(\pi t/13.75)$	0.673	17.79**
Final	$Y_t = 569.986+0.016(t-40.5) + 0.000785(5-40.5)^2$ $- 0.499 \sin(\pi t/13.75) + 0.180 \cos(\pi t/13.75)$ $+ 0.362 \sin(\pi t/20) - 0.022 \cos(\pi t/20)$	0.638	5.23**

** Significant at the 1% level.

Table 7. Central Basin Air Temperature: Monthly Means (°C)

Year	April	May	June	July	Aug.	Sept.	Oct.	Nov.
1967	5.0	9.1	18.6	20.1	20.5	18.1	13.5	7.2
1968	5.6	9.9	16.5	20.0	22.0	20.4	14.5	9.6
1969	5.6	10.1	15.9	21.5	23.1	20.1	14.3	7.2
1970	6.2	11.1	17.3	20.6	22.5	21.5	14.7	9.1
1971	3.6	9.5	18.4	20.8	21.4	20.9	15.7	9.7
1972	4.0	11.0	16.3	19.7	21.6	19.5	13.8	8.2
1973	5.4	10.5	18.1	21.7	23.3	20.0	15.4	8.1
1974	5.2	9.7	15.9	20.6	21.8	17.9	12.8	7.9
1975	3.1	11.6	17.8	23.0	22.0	17.8	14.0	10.1
1976	5.5	10.3	17.8	20.2	20.7	18.1	12.4	4.5
1977	5.4	11.4	16.7	21.8	22.3	20.0	13.2	8.9
1978	3.6	11.7	17.9	21.1	22.8	20.3	14.2	8.5
Mean	4.83	10.49	17.22	20.93	22.00	19.55	14.04	8.25
S.D.	0.98	0.86	1.00	0.94	0.87	1.27	0.98	1.51

Table 8. Residuals and Cumulative Sum Residuals for the Central Basin Air Temperature (°C)

Year		April	May	June	July	Aug.	Sept.	Oct.	Nov.
1967	Residual	0.17	-1.39	1.38	-0.83	-1.5	-1.45	-0.54	-1.05
	CUSUM	0.17	-1.22	0.16	-0.67	-2.17	-3.62	-4.16	-5.21
1968	Residual	0.47	-0.59	-0.72	-0.93	0.0	0.85	0.46	1.35
	CUSUM	0.47	-0.12	-0.84	-1.77	-1.77	-0.92	-0.46	0.89
1969	Residual	0.77	-0.39	-1.32	0.57	1.1	0.55	0.26	-1.05
	CUSUM	0.77	0.38	-0.94	-0.37	0.73	1.28	1.54	0.49
1970	Residual	1.37	0.61	0.08	-0.33	0.50	1.95	0.66	0.85
	CUSUM	1.37	1.98	2.06	1.73	2.23	4.18	4.84	5.69
1971	Residual	-1.23	-0.99	1.18	-0.13	-0.60	1.35	1.66	1.45
	CUSUM	-1.23	-2.22	-1.04	-1.17	-1.77	-0.42	1.24	2.69
1972	Residual	-0.88	0.51	-0.92	-.123	-0.40	-0.05	-0.24	-0.05
	CUSUM	-0.88	-0.37	-1.29	-2.52	-2.92	-2.97	-3.21	-3.26
1973	Residual	0.57	0.01	0.88	0.77	1.30	0.45	1.36	-0.15
	CUSUM	0.57	0.58	1.46	2.23	3.53	3.98	5.34	5.19
1974	Residual	0.37	-0.79	-1.32	-0.33	-0.20	-1.65	-1.24	-0.35
	CUSUM	0.37	-0.42	-1.74	-2.07	-2.27	-3.92	-5.16	-5.51
1975	Residual	-1.73	1.11	0.58	-2.07	0.0	-1.75	-0.04	1.85
	CUSUM	-1.73	-0.62	-0.04	2.03	2.03	0.28	0.24	2.09
1976	Residual	0.67	-0.19	0.58	-0.73	-1.30	-1.45	-1.64	-3.75
	CUSUM	0.67	0.48	1.06	0.33	-0.97	-2.42	-4.06	-7.13
1977	Residual	0.57	0.91	-0.52	0.87	0.30	0.45	-0.84	0.65
	CUSUM	0.57	1.48	0.96	1.83	2.13	2.58	1.74	2.39
1978	Residual	-1.23	1.21	0.68	0.17	0.80	0.75	0.16	0.25
	CUSUM	-1.23	-0.02	0.66	0.83	1.63	2.38	2.54	2.79

Table 9. A Summary of the Air Temperature Pattern for the Central Basin, 1967 to 1978

Year	Spring	Summer	Fall
1967	N	C	C
1968	N	C	W
1969	N	N	N
1970	W	W	W
1971	C	N	W
1972	C	C	C
1973	W	W	W
1974	N	C	C
1975	N	W	N
1976	N	C	C
1977	C	N	N
1978	N	W	N

C - Cold.
N - Normal.
W - Warm.

Table 10. Results of Fitting Model 3.4 to the Surface Temperature Data of Lake Erie, 1967 to 1978

	Western Basin					Central Basin					Eastern Basin				
Year	Maximum temperature $\alpha + R_t$	Phase ϕ	No. of days D_t	R^2	F	Maximum temperature $\alpha + R_t$	Phase ϕ	No. of days D_t	R^2	F	Maximum temperature $\alpha + R_t$	Phase ϕ	No. of days D_t	R^2	F
1967	22.87	-1.768	205.4	0.92	37.9**	22.25	-1.884	206.8	0.76	11.3**	22.82	-1.929	224.0	0.89	24.3**
1968	24.14	-1.877	218.0	0.98	60.7**	22.52	-1.970	228.8	0.99	131.4**	22.07	-2.005	232.8	0.99	109.6**
1969	23.43	-1.907	221.4	0.99	141.3**	22.39	-1.971	229.0	0.98	85.5**	19.78	-1.851	215.0	0.86	14.9**
1970	23.99	-1.873	217.6	0.97	130.0**	22.87	-1.956	227.2	0.96	78.2**	22.14	-1.989	231.0	0.93	45.7**
1971	21.47	-1.749	203.2	0.86	9.0**	19.46	-1.882	218.6	0.81	6.4**	21.02	-1.802	209.2	0.95	18.2**
1972	22.53	-1.880	218.4	0.94	32.3**	21.65	-1.963	228.0	0.96	50.7**	21.31	-1.994	231.6	0.97	62.9**
1973	25.60	-1.809	210.0	0.99	76.4**	23.90	-1.879	218.2	0.99	91.6**	23.78	-1.882	218.6	0.99	809.5**
1974	No Data														
1975	23.09	-1.866	216.8	0.96	47.6**	21.26	-1.937	225.0	0.94	41.6**	21.62	-1.938	225.0	0.93	26.1**
1977	—	—	—			23.17	-1.935	224.8	0.99	280.2**	22.34	-1.970	228.8	0.99	211.4**
1978	—	—	—			22.65	-1.968	228.6	0.97	129.8	22.67	-1.965	228.2	0.98	220.2**

** Significant at the 1% level.

Table 11. The Values of F-Statistic for Testing the Equality of Yearly Water Temperature Cycle

Western Basin		Central Basin		Eastern Basin	
Years	F-statistic	Years	F-statistic	Years	F-statistic
1967-1975	2.77*	1967-1978	3.65**	1967-1978	5.70**

* Significant at 5% level.
** Significant at 1% level.

Table 12. The Mean of the Hypolimnion Temperature, the Temperature at Julian Day 200 and the Slope and Intercept of the Regression Line 1967 to 1980

Year	Mean	Temperature (Julian day 200)	Intercept	Slope
1967	10.92	10.28	10.59	0.0021
1968	13.18	13.31	13.51	-0.0042
1969	11.36	10.96	8.51	0.0489
1970	9.90	9.99	7.31	0.0536
1971	10.96	10.70	9.40	0.0260
1972	10.98	10.36	6.64	0.0743
1973	11.39	10.78	10.36	0.0083
1975	8.17	8.86	5.48	0.0676
1977	11.72	11.78	8.42	0.0672
1978	8.57	8.87	7.31	0.0313
1979	10.15	9.48	7.68	0.0359
1980	11.15	9.31	5.09	0.0843

Table 13. The Estimates of the Parameters of Model 3.4 for Chlorophyll *a* (Western Basin)

Year	Model parameters				Estimated maximum	Coefficient of determination
	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\gamma}_t$	$\hat{\sigma}$		
1968	9.00	-0.708	-1.364	2.283	10.54	0.82
1970	9.77	-1.238	0.736	4.606	11.21	0.74
1972	9.75	3.038	1.935	1.924	13.35	0.77

Table 14. The Values of F_1, F_2 and F_0 for Chlorophyll *a* (Western Basin)

H_1		H_2		H_0	
Degrees of freedom	F_1	Degrees of freedom	F_2	Degrees of freedom	F_0
4,9	1.61	2,13	1.55	6,9	1.30

Table 15. The Estimates of the Parameters of Model 3.6 for Chlorophyll *a* (Western Basin)

$\hat{\alpha}_t$				
1968	1970	1972	$\hat{\beta}$	$\hat{\gamma}$
8.27	10.01	11.81	-0.1585	0.2579

Table 16. The Estimates of the Parameters of Model 3.4 for Chlorophyll *a* (Central Basin)

	Model parameters					Coefficient of
Year	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\sigma}$	Maximum	determination
1968	5.163	-0.751	-1.478	1.299	6.82	0.61
1970	6.425	-0.369	-0.276	2.647	6.89	0.89
1979	3.998	0.666	-0.275	2.589	4.718	0.69
1980	3.268	-0.355	-0.739	0.637	4.087	0.65

Table 17. The Values of F_1, F_2 and F_0 for Chlorophyll *a* (Central Basin)

H_1		H_2		H_0	
Degrees of freedom	F_1	Degrees of freedom	F_2	Degrees of freedom	F_0
6.23	0.266	3.29	4.304*	9.23	1.39

* Significant at the 5% level.

Table 18. The Estimates of the Parameters of Model 3.6 for Chlorophyll a (Central Basin)

$\hat{\alpha}_t$					
1968	1970	1979	1980	$\hat{\beta}$	$\hat{\gamma}$
5.045	6.345	4.327	3.045	0.766	-0.5909

Table 19. The Estimates of the Parameters of Model 3.4 for Chlorophyll a (Eastern Basin)

	Model parameters					
Year	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\gamma}_t$	$\hat{\sigma}$	Maximum	Coefficient of determination
1968	7.962	4.538	-1.952	2.08	8.241	0.55
1970	5.334	-0.031	0.505	1.65	5.808	0.62
1972	3.735	0.153	-0.424	0.885	4.186	0.63
1980	2.698	0.575	0.061	0.652	3.276	0.68

Table 20. The Values of F_1, F_2 and F_0 for Chlorophyll a (Eastern Basin)

H_1		H_2		H_0	
Degrees of freedom	F_1	Degrees of freedom	F_2	Degrees of freedom	F_0
6,16	0.949	3,22	6.175**	9,16	2.492

**Significant at the 1% level.

Table 21. The Estimates of the Parameters of Model 3.6 for Chlorophyll a (Eastern Basin)

$\hat{\alpha}_t$					
1968	1970	1972	1980	$\hat{\beta}$	$\hat{\gamma}$
5.378	5.233	3.675	2.850	0.3582	-0.0670

Table 22. The Estimates of the Parameters of Model 3.4 for TP (Western Basin)

Year	Model parameters $\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\sigma}$	Estimated maximum
1968	0.00922	0.02998	0.028277	0.00255	0.0375
1970	0.04399	-0.00339	0.00468	0.01855	0.0393
1971	0.04521	0.0153	0.02478	0.022102	0.029
1972	0.03137	-0.00264	0.00125	0.00349	0.028
1977	0.0187	0.00243	0.00325	0.00266	0.014
1978	0.0168	-0.000267	0.00244	0.00482	0.014

Table 23. The Values of F_1, F_2 and F_0 for TP (Western Basin)

H_1		H_2		H_1 H_2	
Degrees of freedom	F_1	Degrees of freedom	F_2	Degrees of freedom	F_0
10,20	0.719	5.30	5.475**	15,20	1.778

**Significant at the 1% level.

Table 24. The Estimates of the Parameters of Model 3.6 for TP (Western Basin)

$\hat{\alpha}_t$							
1968	1970	1971	1972	1977	1978	$\hat{\beta}$	$\hat{\gamma}$
0.037	0.045	0.043	0.028	0.0189	0.0193	0.0193	0.0024

Table 25. The Estimates of the Parameters of Model 3.4 for TP (Central Basin)

Year	Model parameters				Minimum
	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\sigma}$	
1968	0.0306	0.0095	-0.0086	0.00242	0.018
1970	0.0231	0.00048	0.0015	0.00975	0.022
1971	0.0209	-0.0021	0.00375	0.005762	0.017
1972	0.0196	-0.0037	-0.00198	0.0046	0.015
1977	0.01367	0.0046	0.00012	0.00169	0.009
1978	0.01189	-0.00101	0.0011	0.00077	0.010
1979	0.01187	0.00154	-0.00039	0.00443	0.010
1980	0.0106	0.00153	-0.00053	0.00157	0.009

Table 26. The Values of F_1, F_2 and F_0 for TP (Central Basin)

H_1		H_2		H_0	
Degrees of freedom	F_1	Degrees of freedom	F_2	Degrees of freedom	F_0
14.24	1.81	1.81	7.233**	21,24	0.354

** Significant at the 1% level.

Table 27. The Estimates of the Parameters of Model 3.6 for TP (Central Basin)

$\hat{\alpha}_t$									
1968	1970	1971	1972	1977	1978	1979	1980	$\hat{\beta}$	$\hat{\gamma}$
0.024	0.023	0.020	0.018	0.014	0.012	0.012	0.011	0.000424	0.00091

Table 28. The Estimates of the Parameters of Model 3.4 for TP (Eastern Basin)

Year	Model parameters $\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\sigma}$	Minimum
1968	0.019	0.00971	-0.00422	0.00237	0.009
1970	0.018	-0.00096	0.00041	0.00649	0.017
1971	0.018	0.00515	0.00231	0.00603	0.012
1972	0.013	0.00201	0.002266	0.00690	0.010
1977	0.013	-0.00107	-0.0013	0.00119	0.011
1978	0.012	-1.283×10^{-6}	-0.00127	0.00157	0.011
1980	0.012	-0.000564	0.00129	0.00248	0.011

Table 29. The Values of F_1, F_2 and F_0 for TP (Eastern Basin)

H_1 Degrees of freedom	F_1	H_2 Degrees of freedom	F_2	H_0 Degrees of freedom	F_0
18,25	0.910	6,37	2.471*	12,25	0.467

* Significant at the 5% level.

Table 30. The Estimates of the Parameters of Model 3.6 for TP (Eastern Basin)

$\hat{\alpha}_t$ 1968	1970	1971	1972	1977	1978	1980	$\hat{\beta}$	$\hat{\gamma}$
0.0118	0.0180	0.0171	0.0130	0.0137	0.0125	0.0119	-0.00052	0.00032

Table 31. The values of F_1, F_2 and F_0 for SRP (Three Basins)

Basin	H_1 d.f.	F_1	H_2 d.f.	F_2	H_0 d.f.	F_0
Western	8,17	2.08	4,25	6.774**	12,17	4.431**
Central	12,30	2.26*	6,42	3.71**	18,30	3.19**
Eastern	8,16	2.23	4,24	3.37*	12,16	3.07*

* Significant at the 5% level.
** Significant at the 1% level.
d.f. - Degrees of freedom.

Table 32. The Estimates of the Parameters of Model 3.6 for SRP (Three Basins)

	$\hat{\alpha}_t$								
Basin	1968	1970	1971	1973	1978	1979	1980	$\hat{\beta}$	$\hat{\gamma}$
Western	0.0121	0.007	0.0029	0.0079	0.00215	---	---	0.00074	-0.000046
Central	0.0059	0.0036	0.00169	0.0049	0.00174	0.00147	0.000522	0.00076	0.00051
Eastern	0.0026	0.0030	0.0029	0.0083	---	0.0014		-0.00071	0.00008

Table 33. The Estimates of α_t and R_t for NO_3NO_2-N (Three Basins)

	Western Basin		Central Basin		Eastern Basin	
Year	$\hat{\alpha}_t$	R_t	$\hat{\alpha}_t$	R_t	$\hat{\alpha}_t$	R_t
1967	0.18	0.13	0.114	0.092	0.079	0.081
1968	0.18	0.15	0.120	0.107	0.100	0.106
1970	0.23	0.18	0.117	0.099	0.119	0.105
1971	0.14	0.14	0.082	0.106	0.091	0.100
1978	0.25	0.21	0.206	0.100	0.180	0.094
1980	--	--	0.212	0.100	--	--

Table 34. The Values of F_2 and F_0 for NO_3NO_2-N (Three Basins)

	H_2		H_0	
Basin	d.f.	F_2	d.f.	F_0
Western	8,20	9.13**	12,20	17.26**
Central	10,26	4.29**	15,26	8.37**
Eastern	9,19	3.89**	12,19	6.77**

** Significant at the 1% level.
d.f. - Degrees of freedom.

Table 35. The Estimates of α_t and R_t for NO_3NO_2-N (Three Basins)

Year	Western Basin		Central Basin		Eastern Basin	
	$\hat{\alpha}_t$	R_t	$\hat{\alpha}_t$	R_t	$\hat{\alpha}_t$	R_t
1967, 1968 and 1970	0.217	0.178	0.118	0.099	0.111	0.105
1971	0.138	0.140	0.082	0.106	0.091	0.100
1978 and 1980	--	--	0.209	0.101	--	--

Table 36. The Values of F_1, F_2 and F_0 for SRP (Three Basins)

Basin	H_1		H_2		H_0	
	d.f.	F_1	d.f.	F_2	d.f.	F_0
Western	6,16	0.48	3,22	3.07*	9,16	1.20**
Central	8,22	3.00*	4,30	57.23**	12,22	31.26**
Eastern	6,15	2.17	3,21	5.79**	9,15	4.02**

* Significant at the 5% level.
** Significant at the 1% level.

Table 37. The Estimates of the Parameters of Model 3.6 for SRP (Three Basins)

Basin	$\hat{\alpha}_t$					$\hat{\beta}_t$	$\hat{\gamma}_t$
	1968	1970	1971	1978	1980		
Western	0.023	0.027	0.017	0.008	--	0.005	0.001
Central	0.027	0.030	0.020	0.011	0.017	0.007	0.005
Eastern	0.023	0.027	0.027	0.008	--	-0.0053	0.0012

Table 38. Results of Fitting Model 3.4 to Log Secchi Disk Depth of Lake Erie, 1967 to 1978

	Western Basin					Central Basin					Eastern Basin				
Year	Amplitude $\alpha + R$	Phase ϕ	No. of Days D_t	R^2	F	Amplitude $\alpha + R$	Phase ϕ	No. of Days D_t	R^2	F	Amplitude $\alpha + R$	Phase ϕ	No. of Days D_t	R^2	F
1967	3.27	0.32	36.75	0.71	8.44*	5.49	1.77	206.1	0.86	22.05**	7.84	1.89	219.51	0.91	23.34**
1968	1.82	1.68	195.59	0.50	0.986	3.17	1.76	204.90	0.76	4.86*	4.20	2.03	235.38	0.82	4.49**
1969	2.25	1.83	213.15	0.42	1.09	3.40	1.80	208.71	0.43	1.48	4.18	1.70	197.64	0.48	2.28*
1970	1.68	1.62	188.09	0.49	3.30*	4.53	1.84	213.21	0.86	22.29**	4.95	1.88	218.07	0.79	13.07**
1971	—	3.91	—	0.20	0.37	4.20	1.61	187.20	0.55	1.80	5.72	1.58	183.59	0.80	3.90**
1972	2.57	1.76	204.32	0.45	0.82	4.19	1.72	199.27	0.67	4.05*	3.98	2.01	233.41	0.65	2.83*
1973						3.74	1.85	214.73	0.99	193.2**	4.27	1.81	209.86	0.82	4.53*
1975	2.03	1.77	205.36	0.40	1.34	5.14	1.82	210.97	0.87	16.77*	4.89	1.88	218.7	0.85	11.56**
1977						5.00	1.92	222.25	0.63	4.20**	4.87	2.07	240.22	0.76	7.86**
1978						5.62	1.50	173.69	0.33	1.99*	5.13	1.80	208.59	0.23	1.04
All years	1.89	1.67	193.59	0.27	9.63**	4.37	1.68	195.49	0.48	34.17**	4.61	1.78	207.35	0.39	20.85**

* Significant at the 5% level.
** Significant at the 1% level.

Table 39. The Values of F-Statistic for Testing the Equality of the Yearly Secchi Disk Seasonal Cycle

Western Basin		Central Basin		Eastern Basin	
Years	F-Statistic	Years	F-Statistic	Years	F-Statistic
1967-1975	0.984	1967-1978	2.384**	1967-1978	3.475**

**Significant at the 1% level.

Table 40. Results of Fitting Model 3.4 to Log Turbidity Data of Lake Erie, 1967 to 1972

	Western Basin					Central Basin					Eastern Basin				
Year	Amplitude $\alpha_{1t} + R_{1t}$	Phase ϕ_{1t}	No. of days D_{1t}	R^2	F	Amplitude $\alpha_{1t} + R_{1t}$	Phase ϕ_{1t}	No. of days D_{1t}	R^2	F	Amplitude $\alpha_{1t} + R_{1t}$	Phase ϕ_{1t}	No. of days D_{1t}	R^2	F
1967	---	--	—	—	—	-0.095	1.36	206.43	0.78	12.72**	-0.51	1.27	217.11	0.94	40.55**
1968	-0.07	1.35	208.20	0.98	83.92**	-1.68	1.25	219.72	0.81	6.4*	-1.66	1.08	239.74	0.94	14.87**
1969	-1.58	1.39	199.64	0.917	22.18**	-1.58	1.97	136.65	0.71	5.02*	-1.59	1.47	194.94	0.94	40.07**
1970	1.00	1.35	208.03	0.699	6.97*	-0.38	1.36	207.09	0.86	18.18**	-0.566	1.52	188.2	0.74	8.55**
1971	0.77	1.56	183.61	0.98	44.7**	-0.59	1.43	198.32	0.96	23.59**	-0.92	1.43	198.67	0.99	2627.7**
1972	0.89	1.31	212.5	0.62	3.23**	-0.19	1.29	215.70	0.45	1.64	-0.15	1.01	247.1	0.21	0.54**
All years		1.47	194.49	0.59	28.12**	-0.56	1.37	205.75	0.53	23.56**	-0.69	1.37	206.15	0.50	18.68**

* Significant at the 5% level.
** Significant at the 1% level.

Table 41. The Values of F-Statistic for Testing the Equality of the Yearly Turbidity Seasonal Cycle

Western Basin		Central Basin		Eastern Basin	
Years	F-Statistic	Years	F-Statistic	Years	F-Statistic
1967-1972	2.970**	1967-1972	8.956**	1967-1972	2.905**

** Significant at the 1% level.

Table 42. The Ratio of Ammonia to Nitrite and Nitrate

Year	Western Basin	Central Basin	Eastern Basin
1967	0.106	0.229	0.207
1970	0.124	0.254	0.243
1971	0.123	0.244	—
1978	—	0.053	—
1980	—	0.081	--

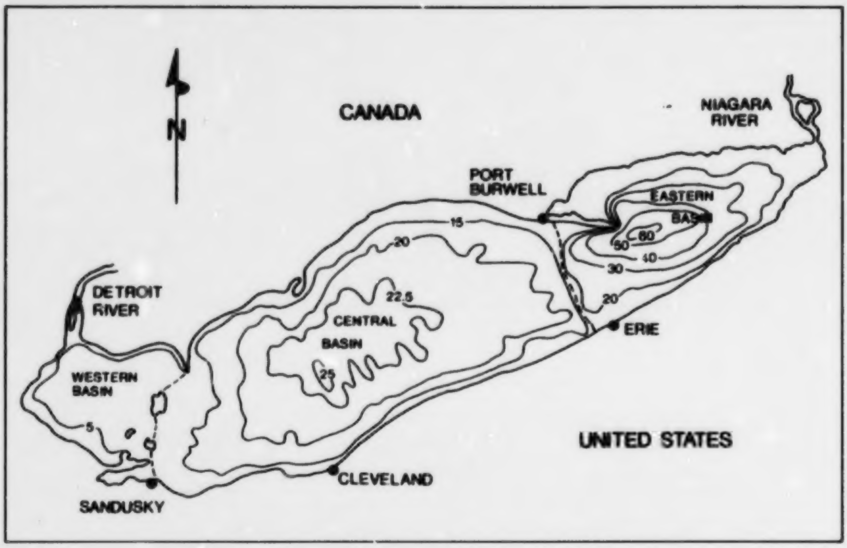

CANADA
NIAGARA RIVER
N
PORT BURWELL
EASTERN BASIN
15
20
60
50
40
30
DETROIT RIVER
22.5
CENTRAL BASIN
ERIE
WESTERN BASIN
25
5
UNITED STATES
CLEVELAND
SANDUSKY

Figure 6. The boundaries of the Western, Central and Eastern basins of Lake Erie.

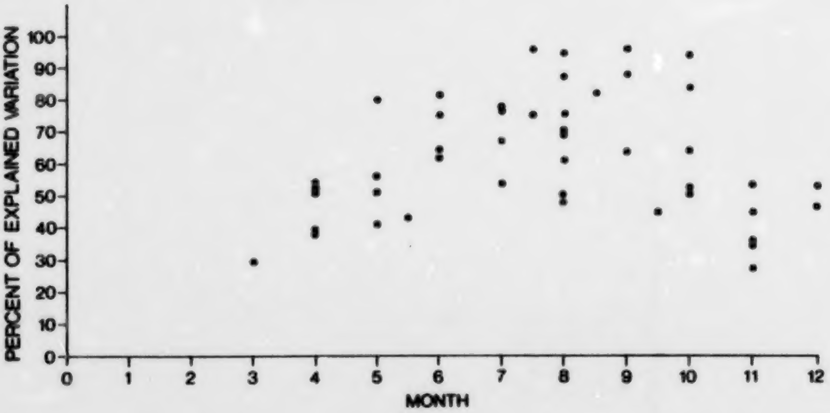


Figure 7. The values of EC for each month using temperature (1968–1976).

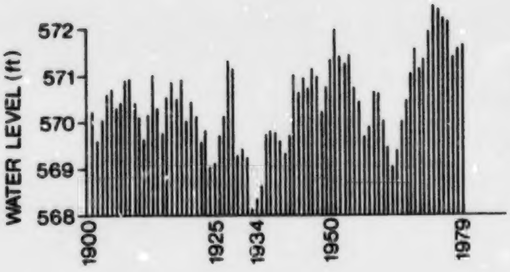


Figure 8. Annual mean water level for the Central Basin.

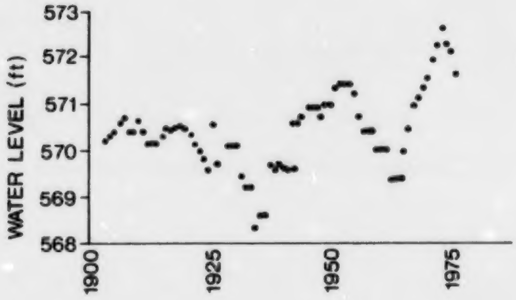


Figure 9. Five-year moving median.

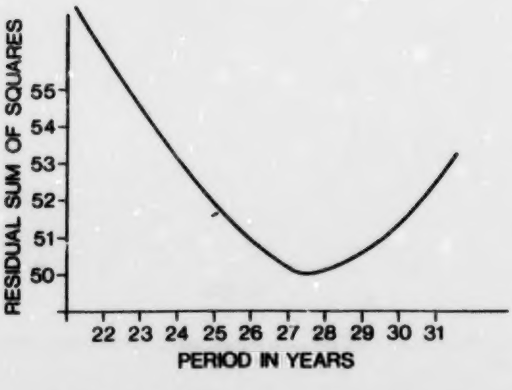


Figure 10. Residual sum of squares against years.

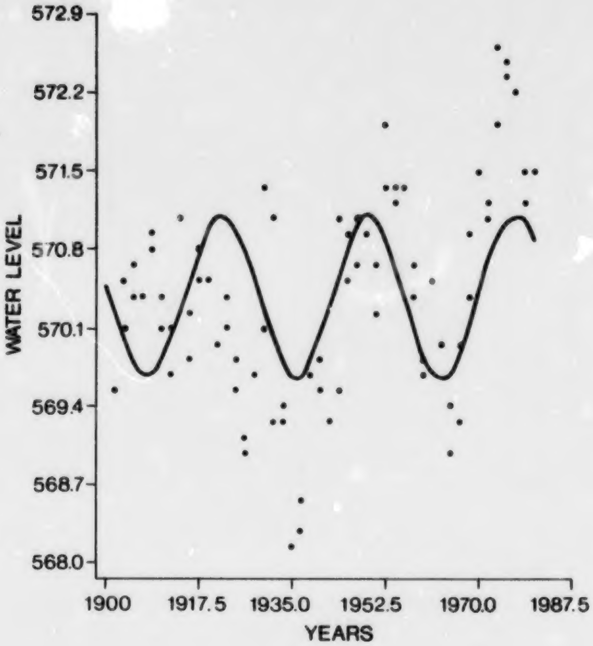


Figure 11. The observed water level and the estimated water level against years.

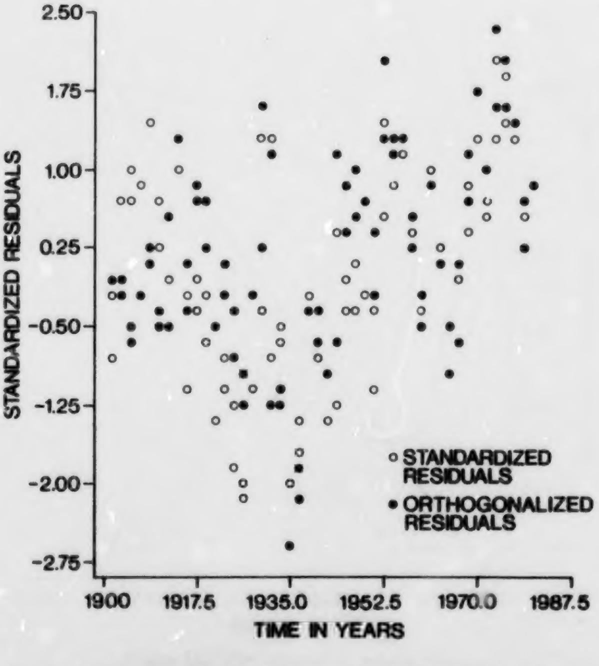


Figure 12. The standardized residuals and orthogonalized residuals against years.

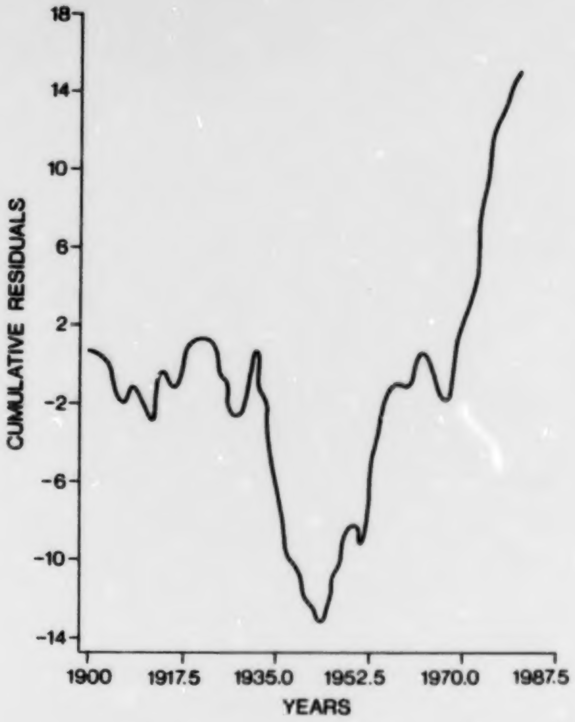


Figure 13. Cumulative standardized residuals against years.

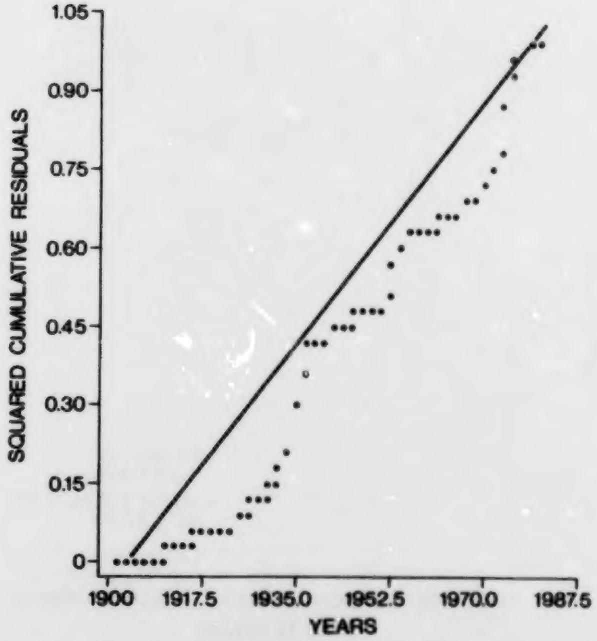


Figure 14. Cumulative squared residuals and their expected values.

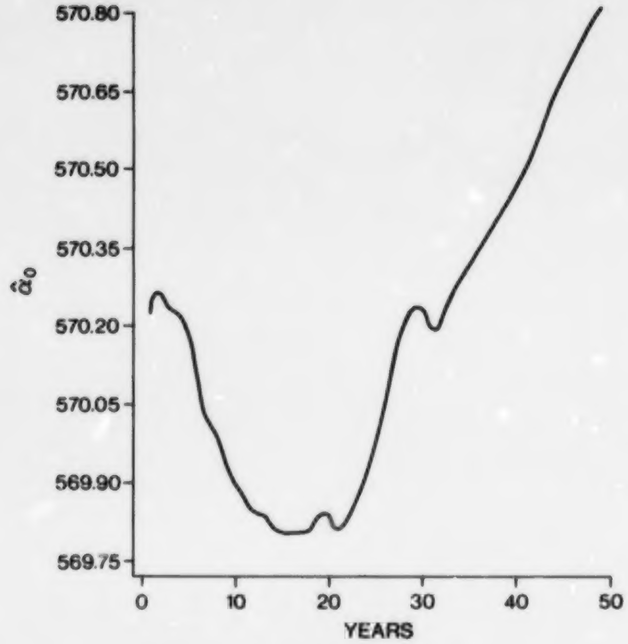


Figure 15. The values of $\hat{\alpha}_0$ against years.

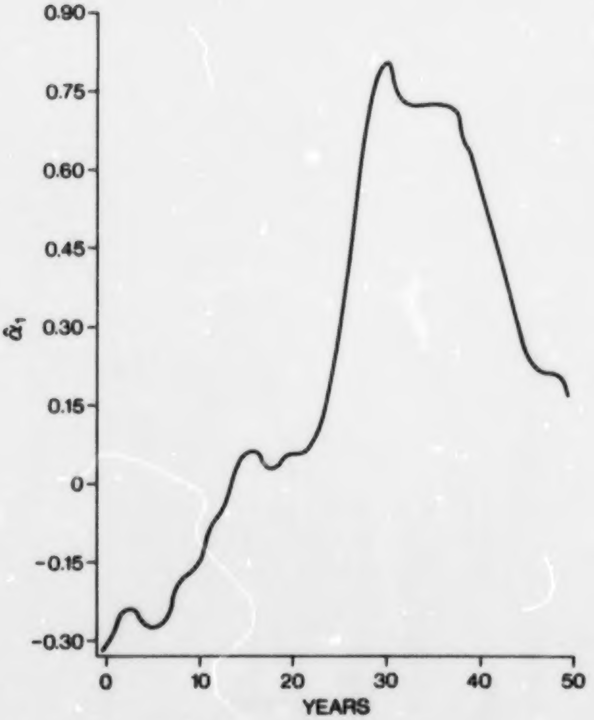


Figure 16. The values of $\hat{\alpha}_1$ against years.

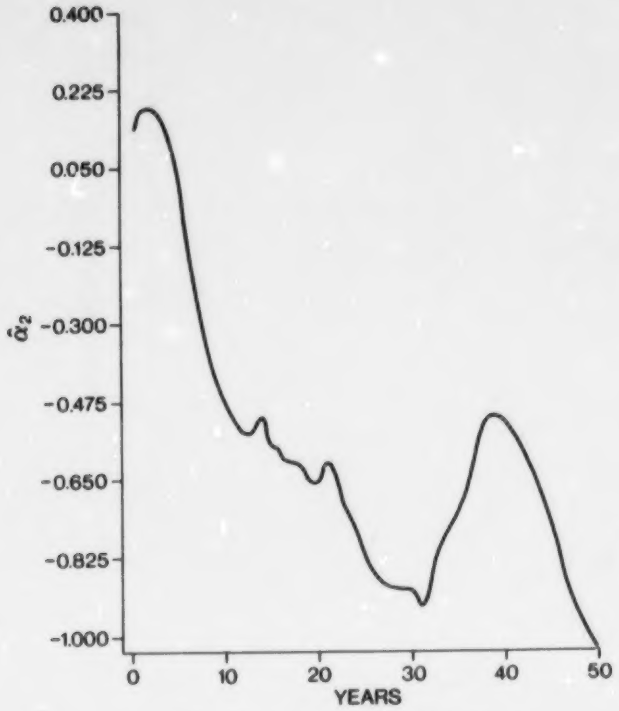


Figure 17. The values of $\hat{\alpha}_2$ against years.

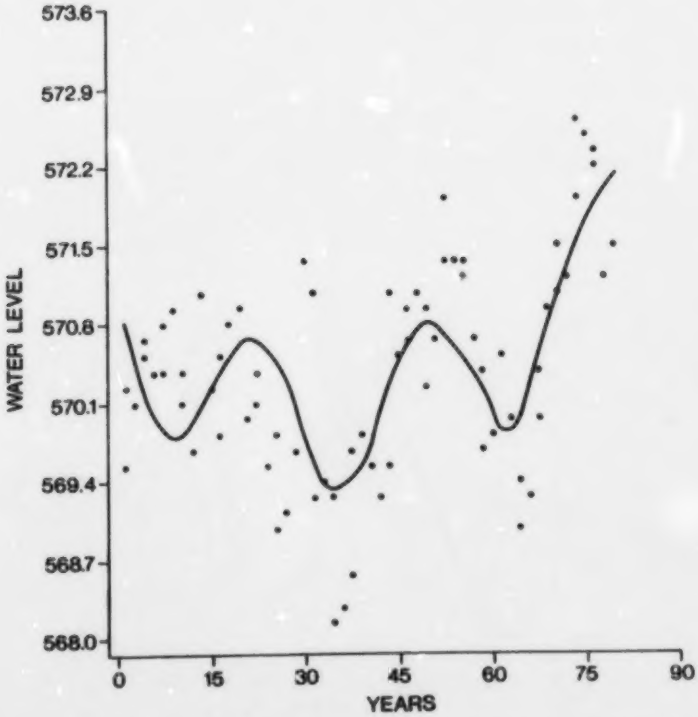


Figure 18. The observed and estimated water level data using model 3.3.

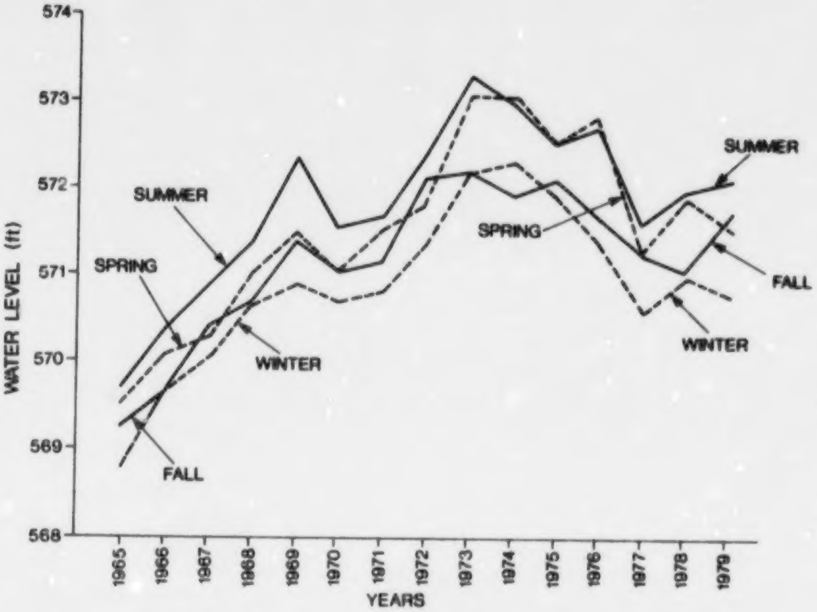


Figure 19. Spring, summer, fall and winter mean water level.

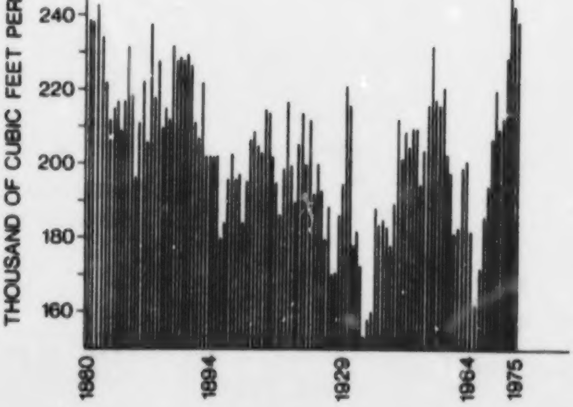


Figure 20. Mean yearly Niagara River flows at Queenston.

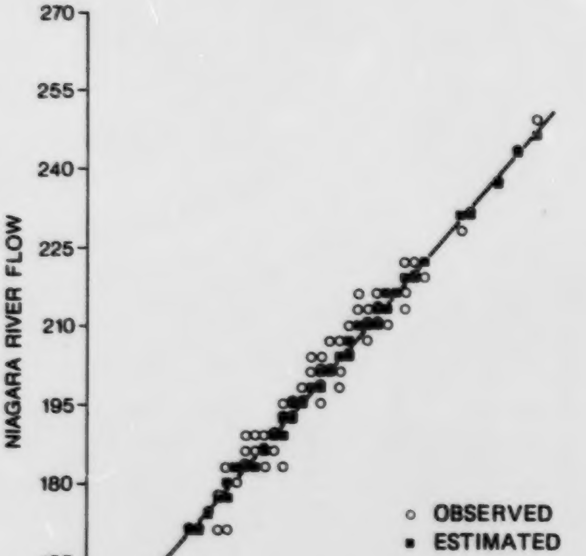


Figure 21. The relationship between the Niagara River flow and Lake Erie water level.

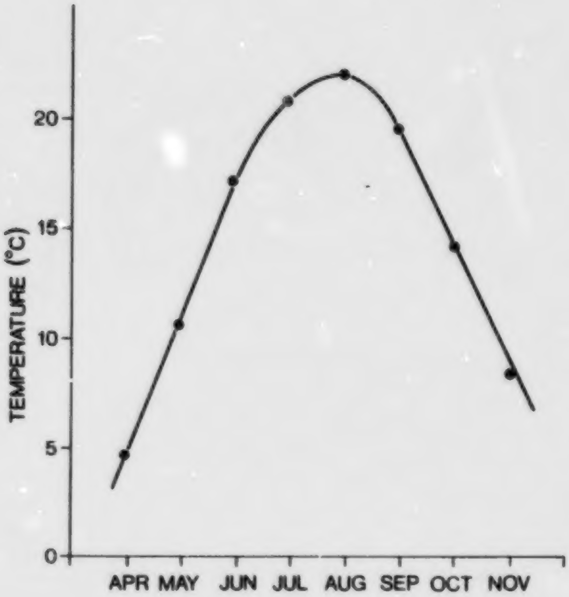


Figure 22a. Monthly air temperature means for the years 1967–1978 (Central Basin).

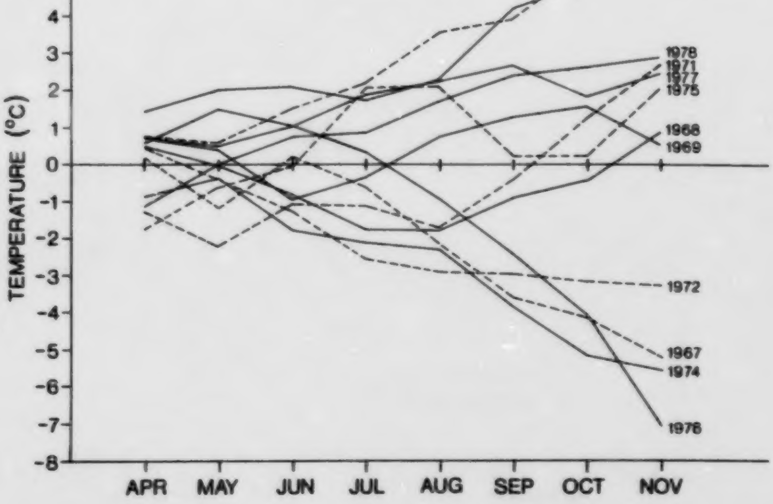


Figure 22b. Plots of the CUSUM for the air temperature for the years 1967–1978 (Central Basin).

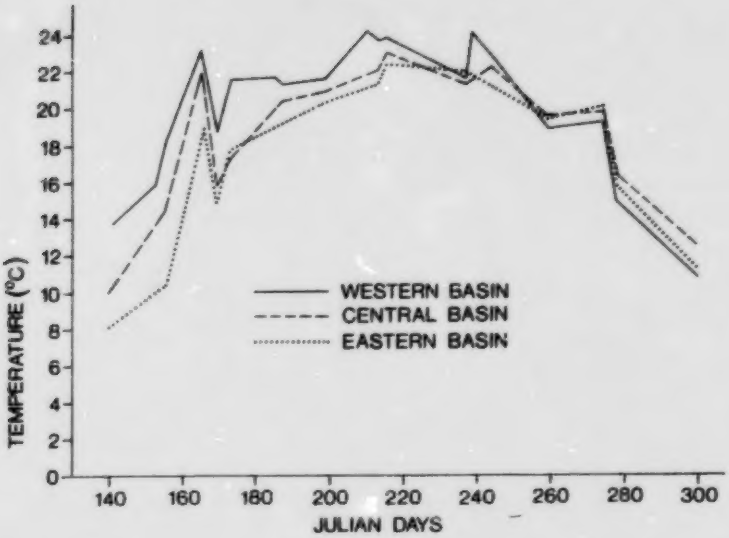


Figure 23. The seasonal temperature cycles for the Western, Central and Eastern basins.

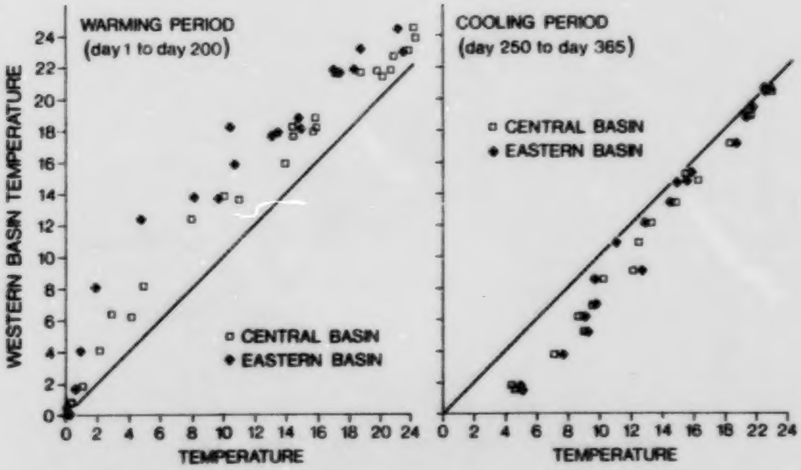


Figure 24. The mean of surface water temperature for the Western Basin against that for the Central and Eastern basins and for the warming and cooling periods.

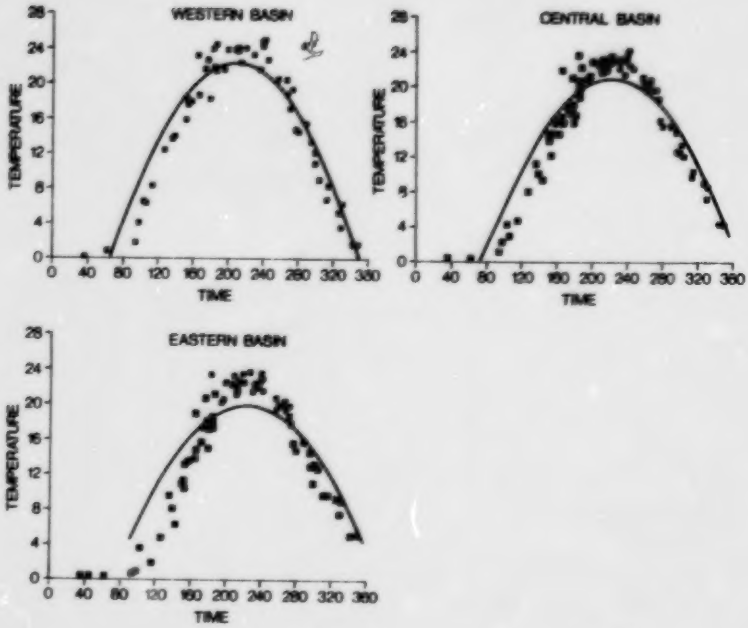


Figure 25. The mean surface water temperatures and their estimated values from model 3.4.

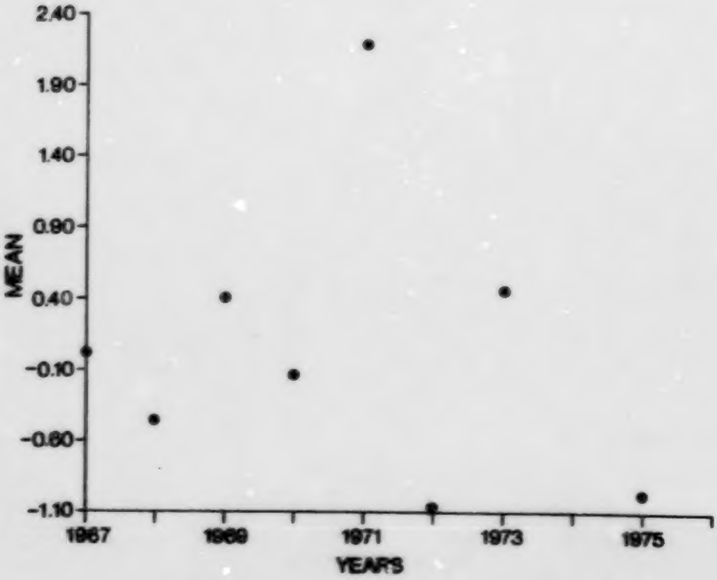


Figure 26a. Mean residual temperature for the Western Basin.

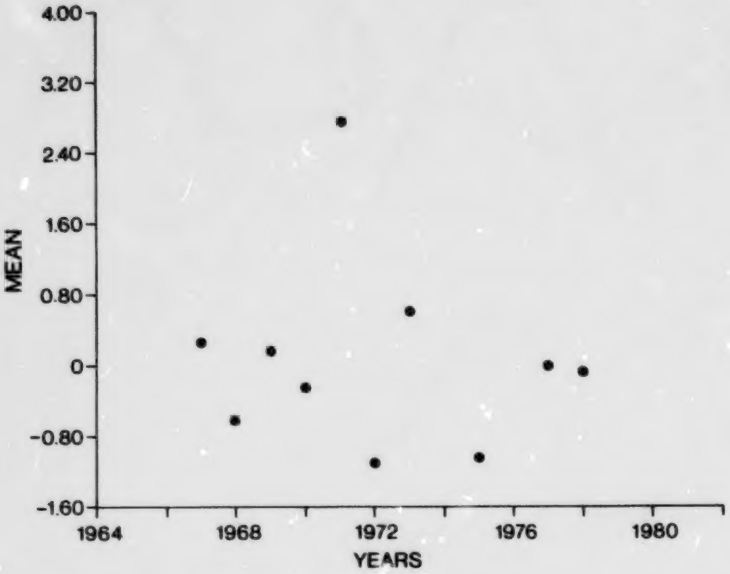


Figure 26b. Mean residual temperature for the Central Basin.

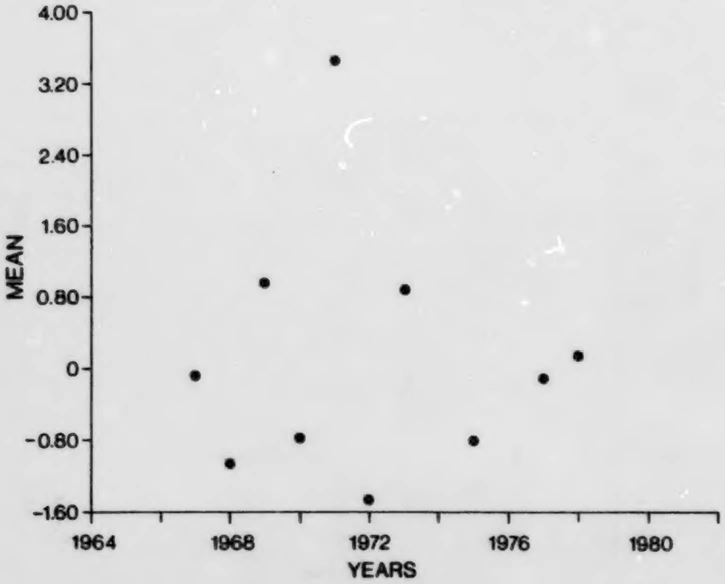


Figure 26c. Mean residual temperature for the Eastern Basin.

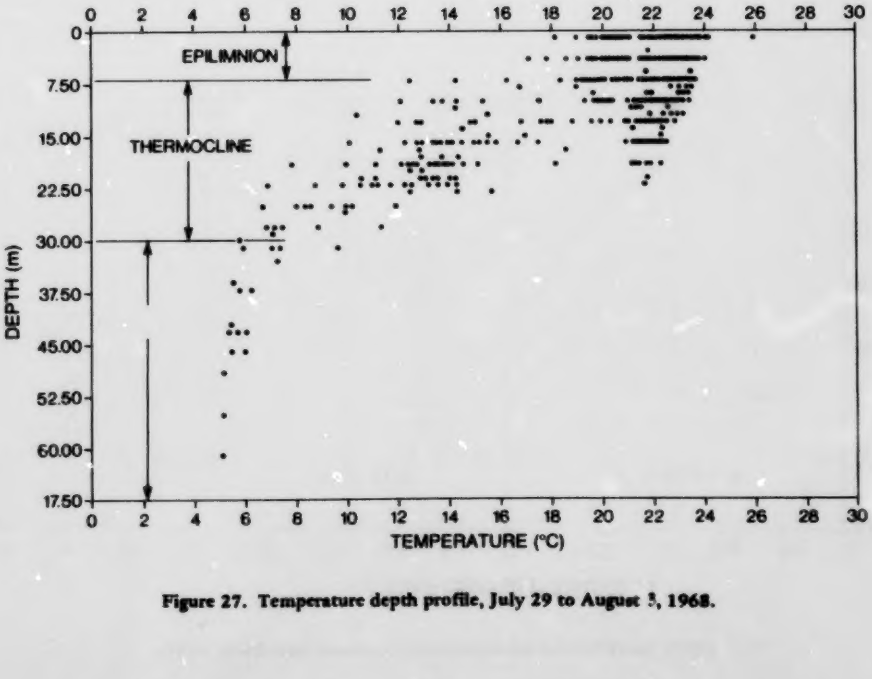


Figure 27. Temperature depth profile, July 29 to August 3, 1968.

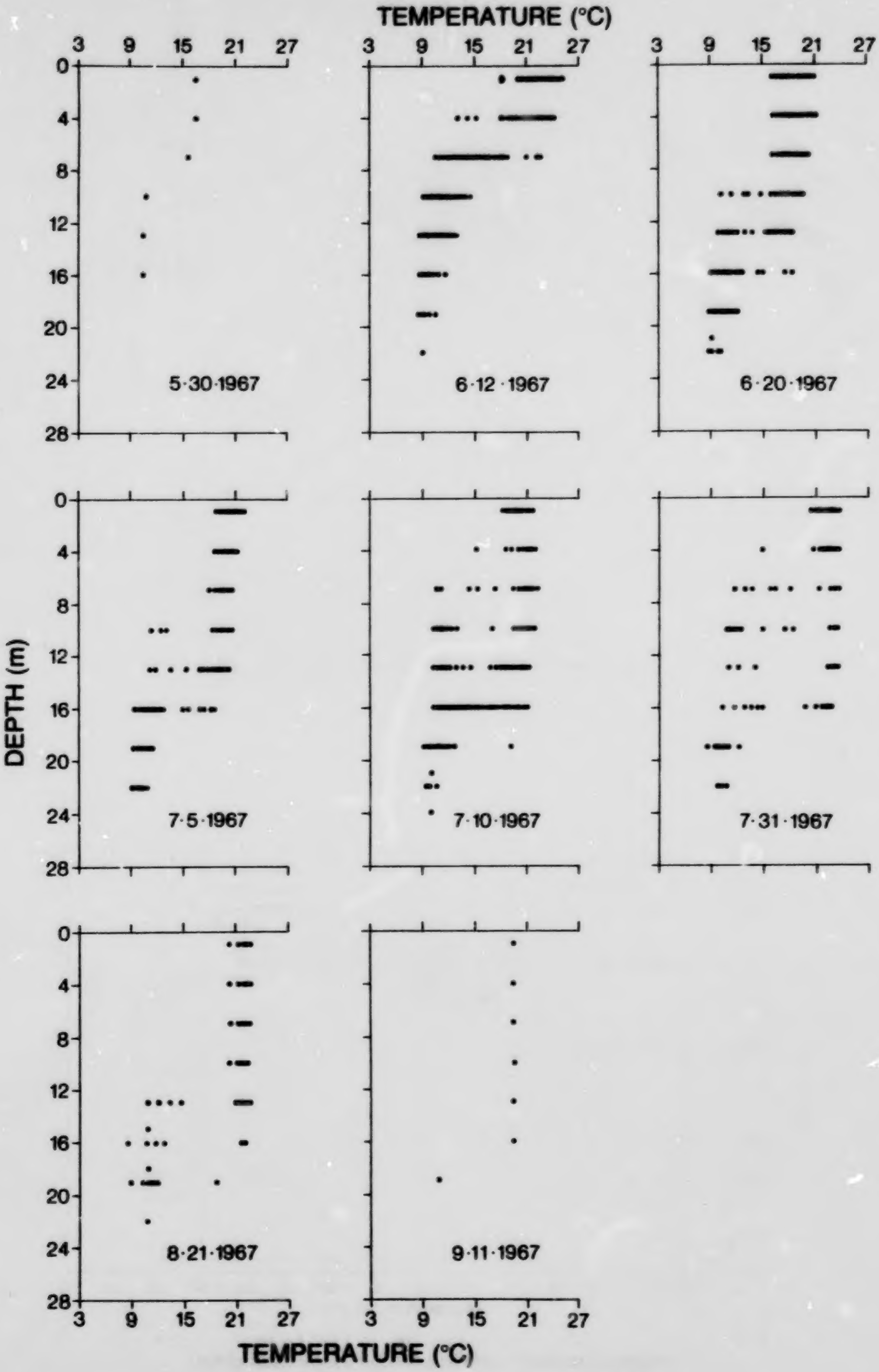


Figure 28. Temperature depth profile for the Central Basin, 1967.

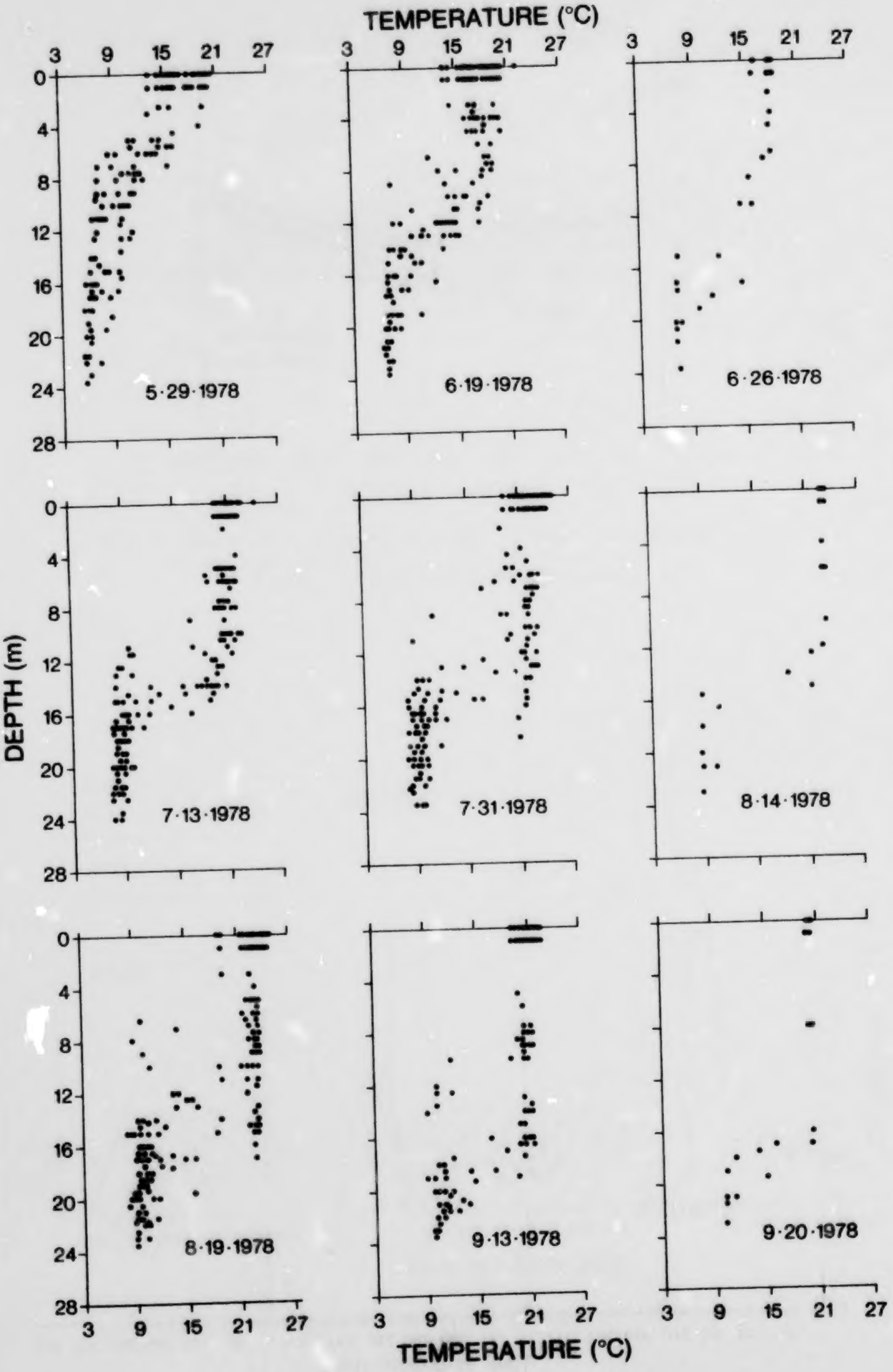


Figure 29. Temperature depth profile for the Central Basin, 1978.

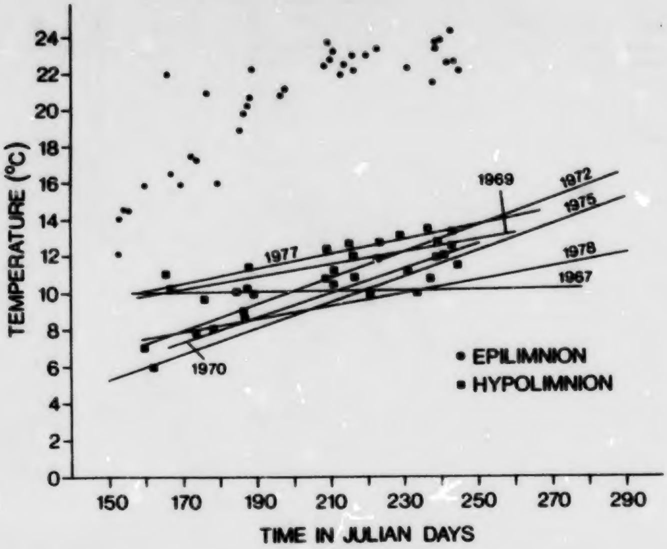


Figure 30. Mean temperature values for the epilimnion and hypolimnion for the Central Basin during 1967–1978 and the fitted regression equations for the hypolimnion data.

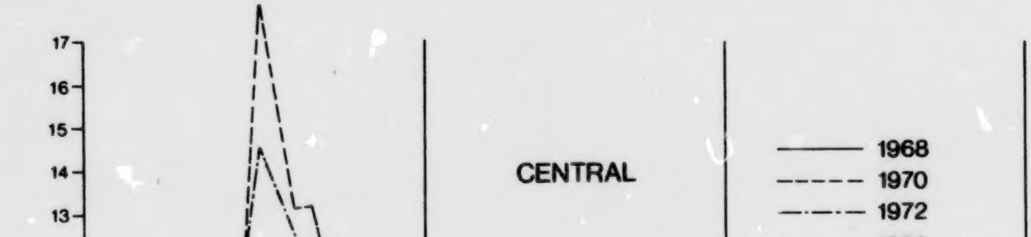


Figure 31. The mean uncorrected chlorophyll a against time.

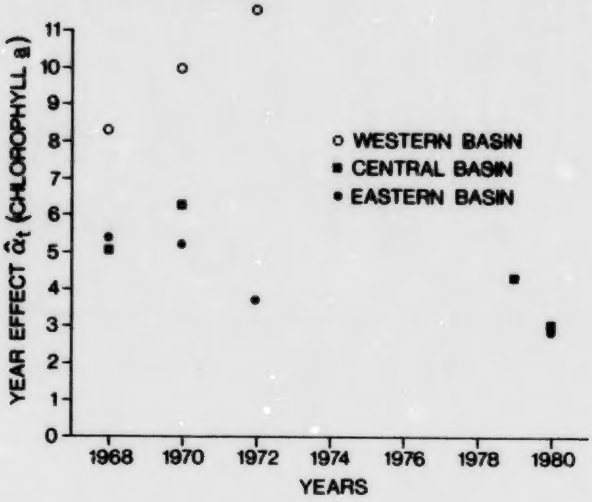


Figure 32. The estimates of the year effects against years (chlorophyll a).

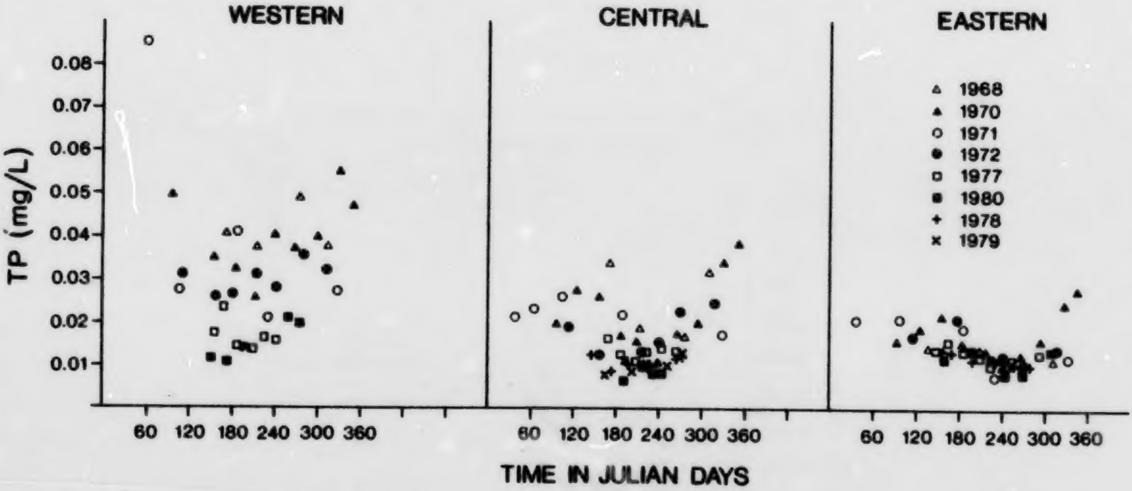


Figure 33. Mean TP against Julian days for the Western, Central and Eastern basins.

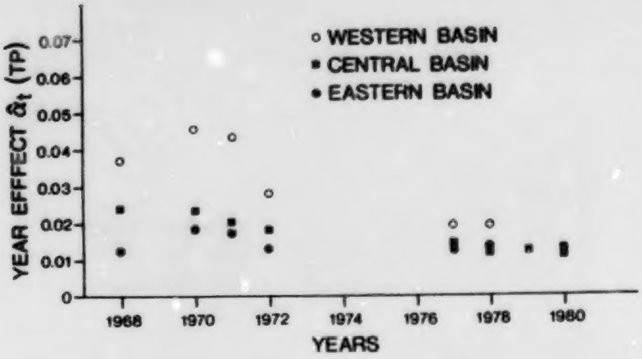


Figure 34. The estimates of the year effects against years (TP).

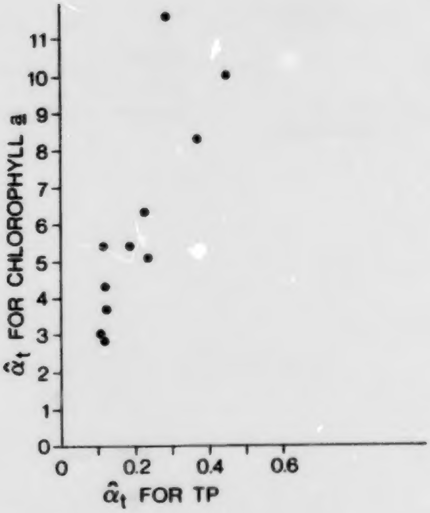


Figure 35. The plot of the year effects for chlorophyll a against those for TP.

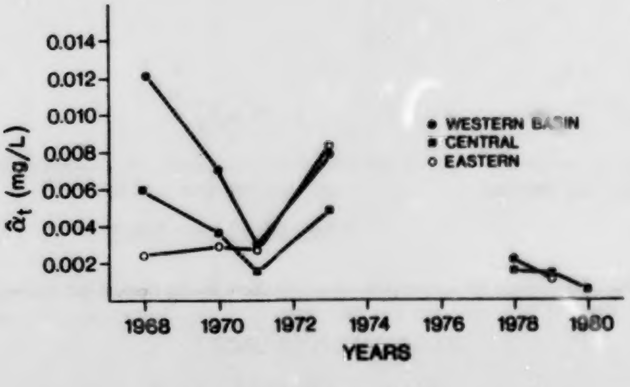


Figure 36. Time trend for SRP.

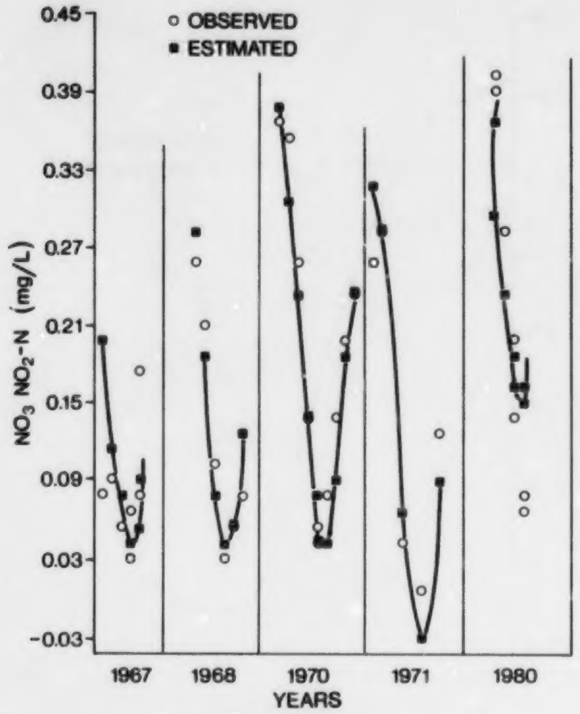


Figure 37. Observed and estimated values for NO_3NO_2-N in the Western Basin.

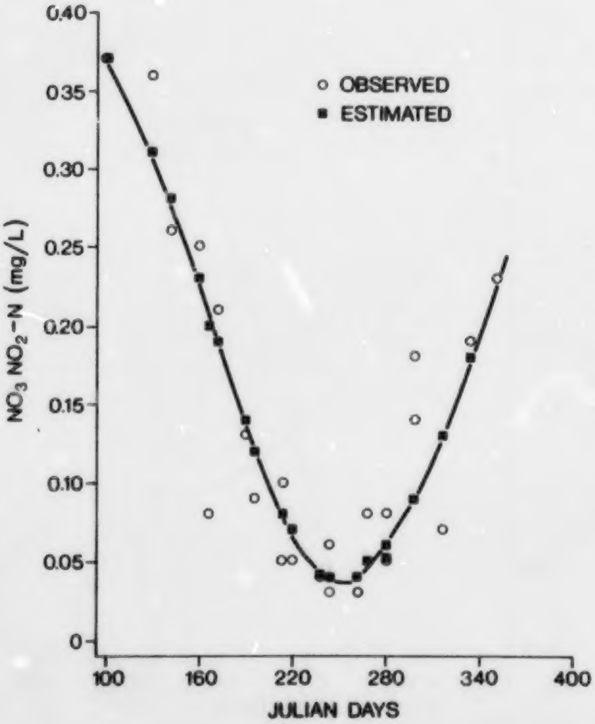


Figure 38a. Observed and estimated values for NO_3NO_2-N in the Western Basin for 1967, 1968 and 1970.

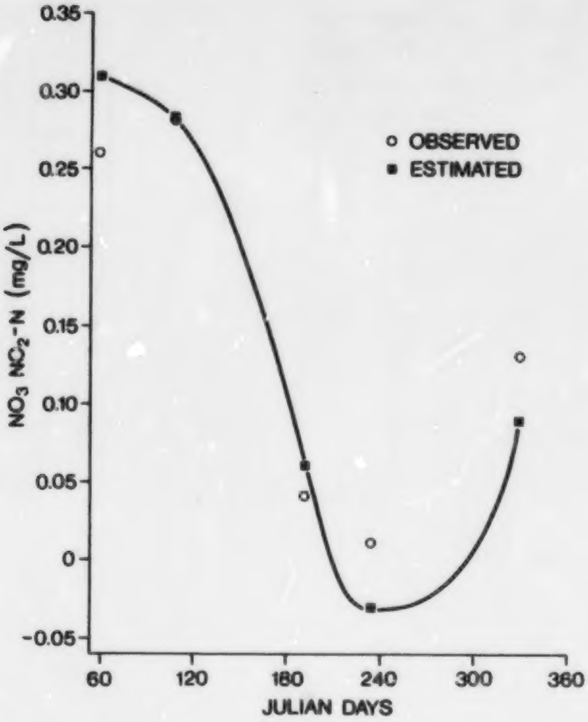


Figure 38b. Observed and estimated values for NO_3NO_2-N in the Western Basin for 1971.

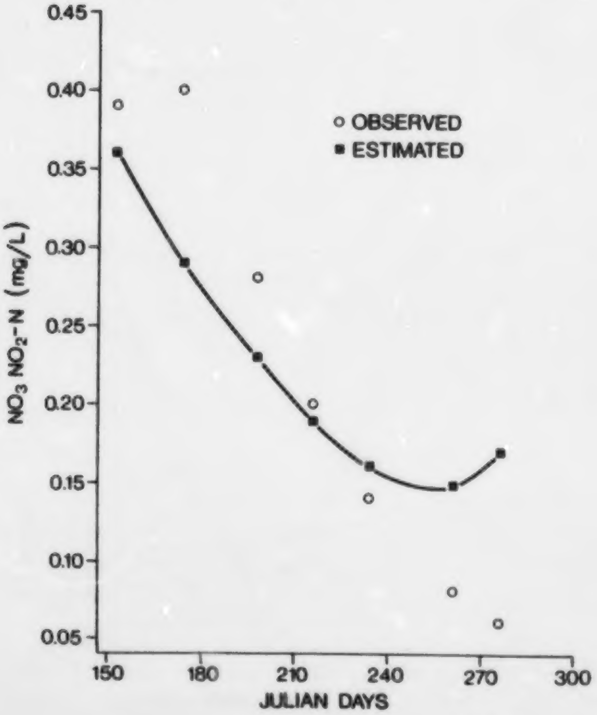


Figure 38c. Observed and estimated values for NO_3NO_2-N in the Western Basin for 1978.

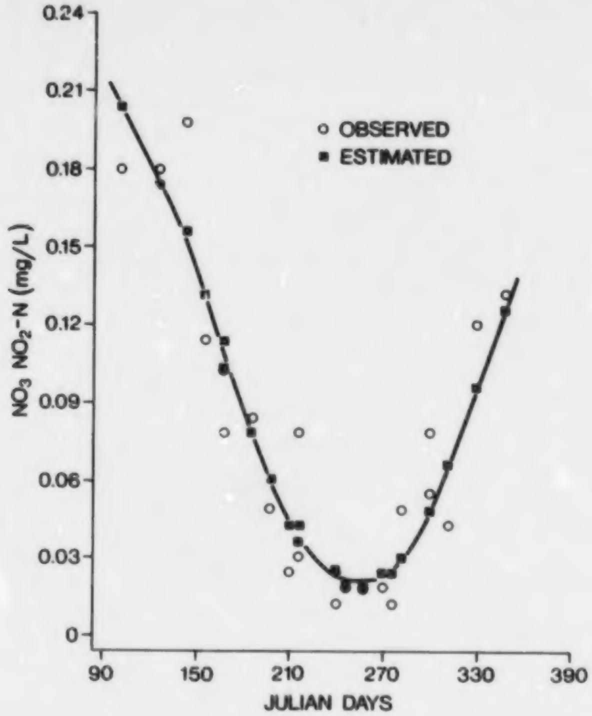


Figure 39a. Observed and estimated values for NO_3NO_2-N in the Central Basin for 1967, 1968 and 1970.

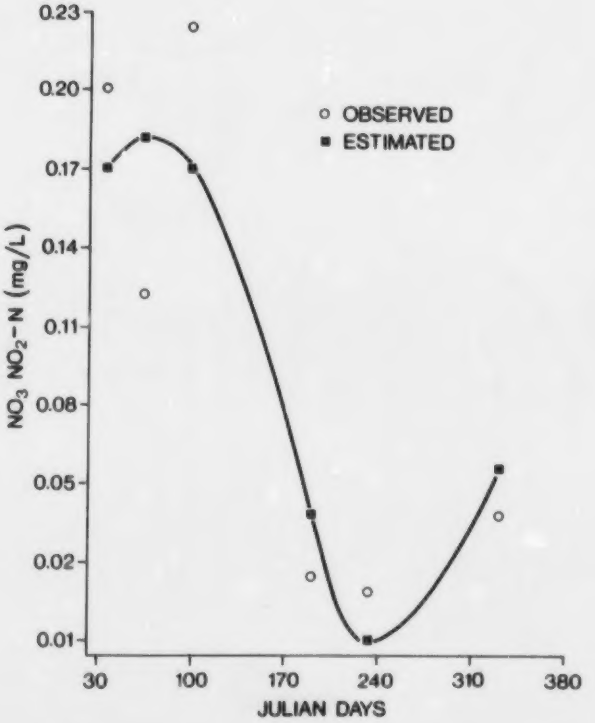


Figure 39b. Observed and estimated values for NO_3NO_2-N in the Central Basin for 1971.

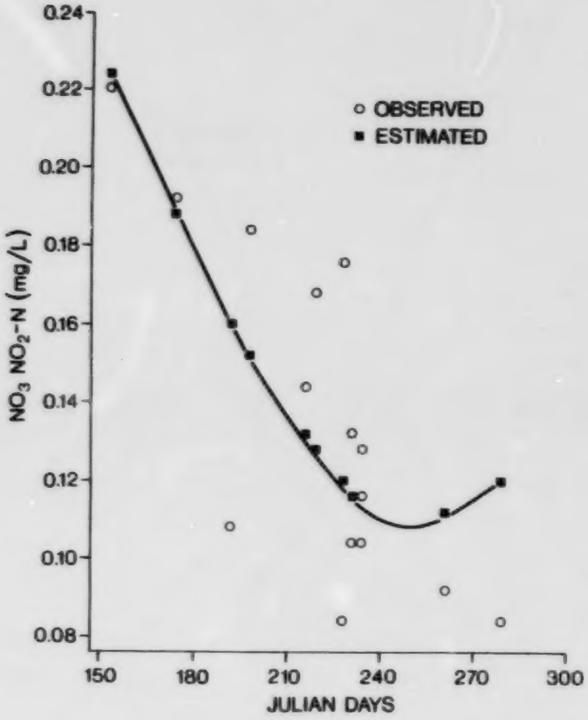


Figure 39c. Observed and estimated values for NO_3NO_2-N in the Central basin for 1978.

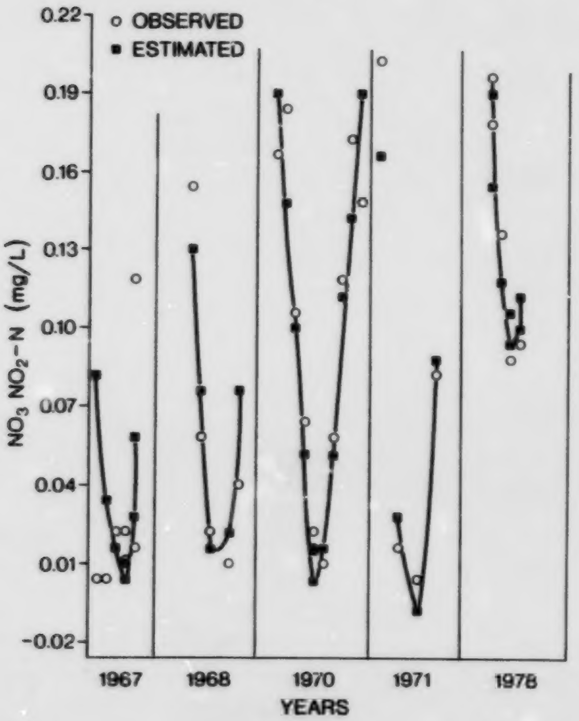


Figure 40. Observed and estimated values of NO_3NO_2-N in the Eastern Basin.

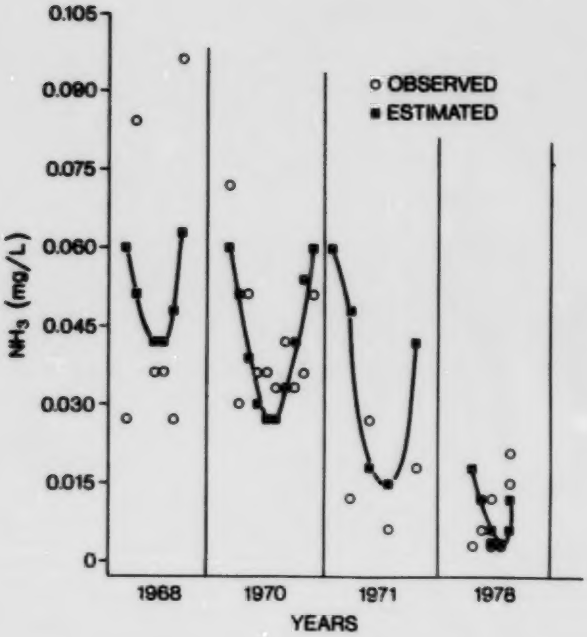


Figure 41. Observed and estimated NH_3 values in the Western Basin.

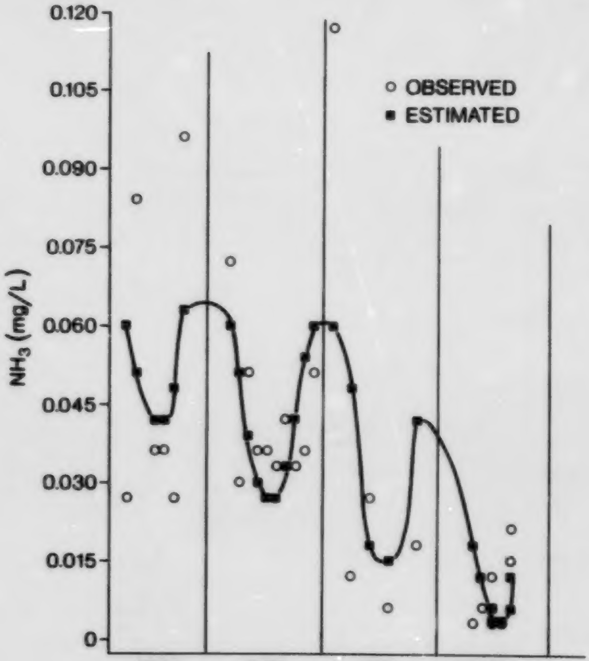


Figure 42. Observed and estimated NH_3 values in the Central Basin.

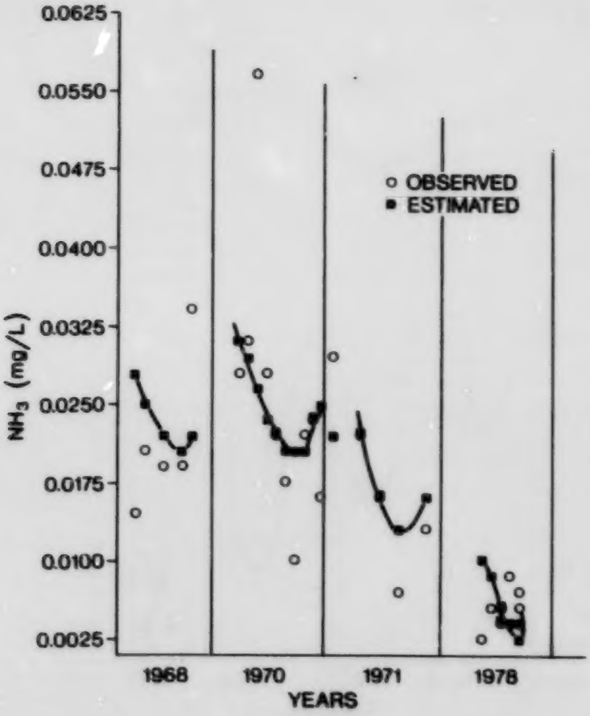


Figure 43. Observed and estimated NH_3 values in the Eastern Basin.

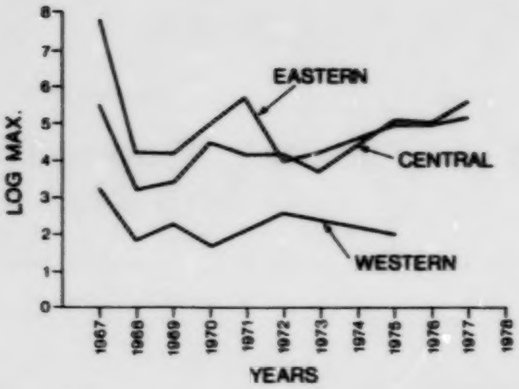


Figure 44. The estimated log Secchi disk value for each basin against years.

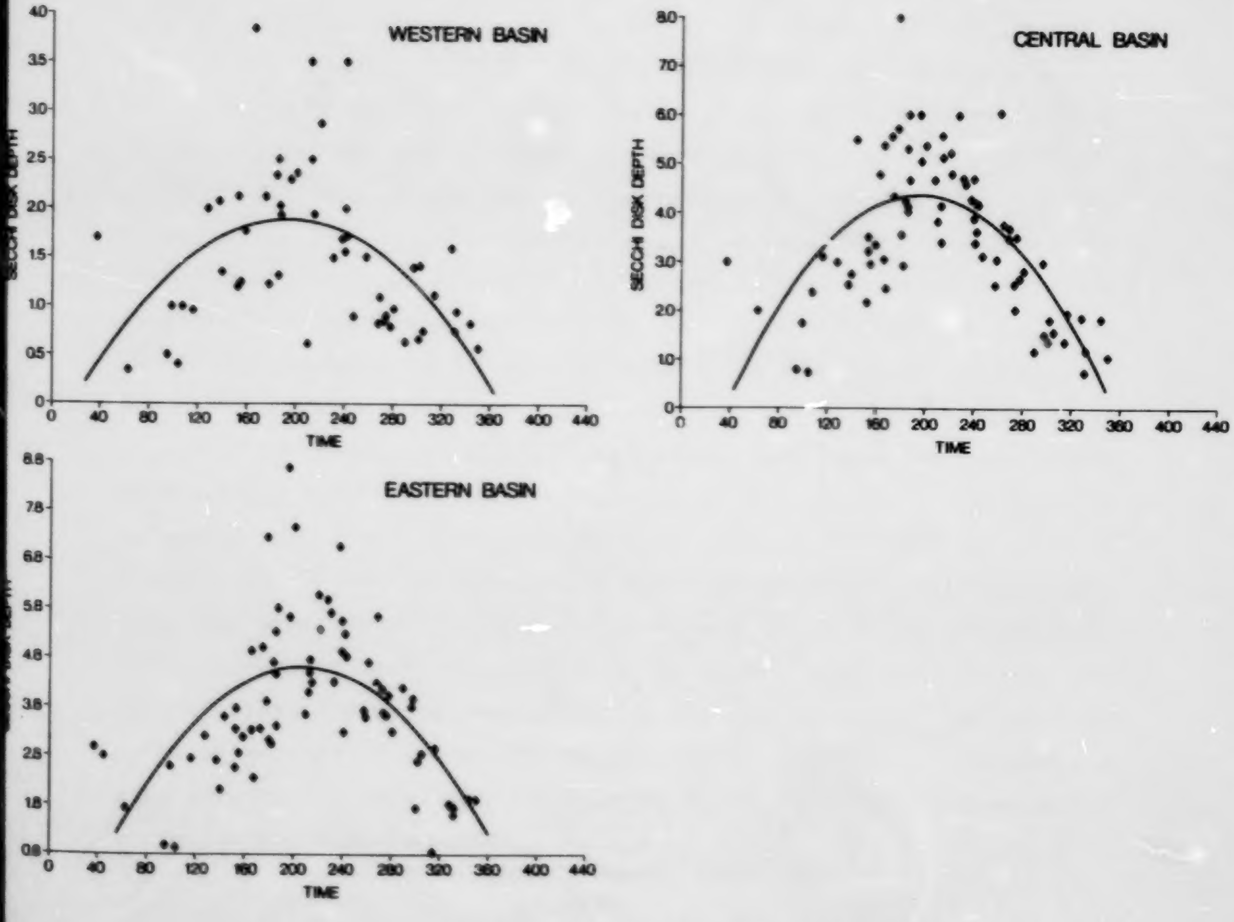


Figure 45. The observed Secchi disk depths and their estimated values from model 3.4.

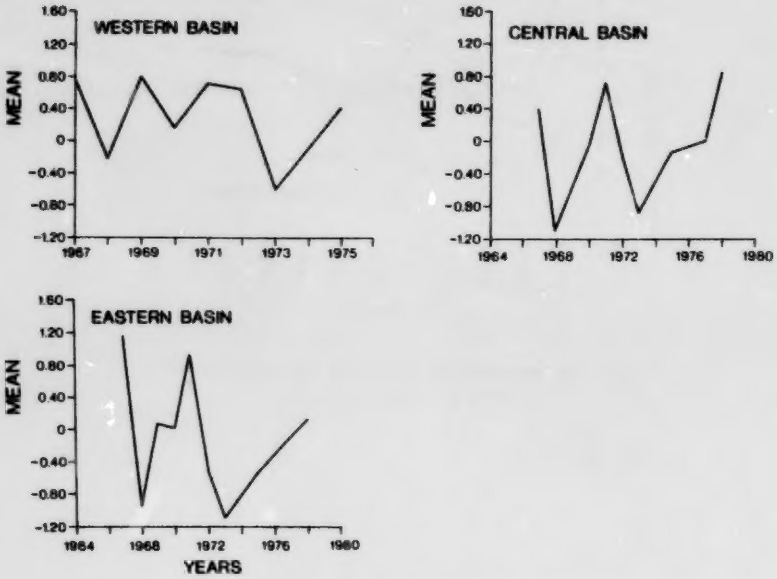


Figure 46. Average residual for Secchi disk.

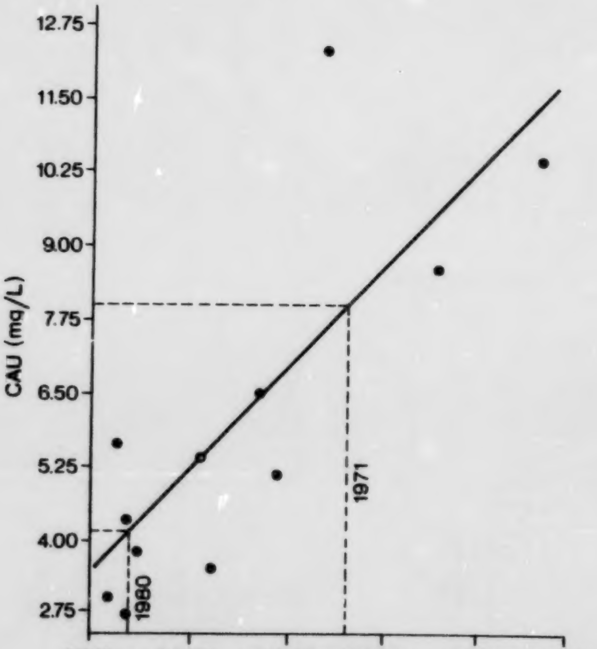


Figure 47. Relationship between chlorophyll a year effect and TP year effect. CAU – Uncorrected chlorophyll a year effect values.

CHAPTER 4

Spatial and Temporal Variability of Dissolved Oxygen in Lake Erie

by J.E. Anderson, A.H. El-Shaarawi, S.R. Esterby and T.E. Unny

ABSTRACT

A non-hierarchical nearest-centroid clustering method was used to separate data pairs consisting of the dissolved oxygen concentration and temperature into four groups corresponding to hypolimnetic water of the Central and Eastern basins and non-hypolimnetic water of these basins. For the stations that were common to all cruises within a year and were classified as being in the hypolimnion, initial dissolved oxygen concentration and depletion rates were calculated and tests about their constancy were performed using weighted regression analysis and regression models with the time structure of the data explicitly incorporated in the models. The yearly uncorrected depletion rates for 1967 to 1980 were similar to values previously reported by several authors, indicating that this semi-objective clustering procedure provides a practical alternative to subjective selection of data. The conclusions about constancy of initial concentrations and depletion rates based on an unweighted regression analysis were shown to differ from those of weighted regression. Using regression with empirical weights, it was found that neither the initial dissolved oxygen concentration nor the depletion rate remained constant between 1967 and 1980 in the Central Basin but the depletion rate remained constant and the initial dissolved oxygen concentration varied in the Eastern Basin.

INTRODUCTION

The dissolved oxygen concentration in a lake serves as a indicator of water quality and is useful for the purpose of lake classification because of its role as a regulator of metabolic processes. The depletion of hypolimnetic oxygen has long been recognized as a natural occurrence which begins each spring and continues until late fall in the Central and Eastern basins of Lake Erie. The thermal gradient usually forms about early June, marking the beginning of the summer stagnation period. The stratification period in Lake Erie has an approximate duration of 110 days in the Central Basin and 140-175 days in the Eastern Basin. There is normally no evidence of persistent stratification in the Western Basin. Oxygen depletion in the hypolimnion zone is characteristic of the summer stratification period. It can be attributed primarily to two factors. First, oxygen replenishment in the hypolimnion is blocked by the temperature-density gradient of the thermocline. Secondly, the decay of plants and animals causes additional organic material to settle in the hypolimnion zone. Since bacteria inhabit the bottom sediment and they require oxygen for their metabolic processes to decompose the organic material aerobically, there is an undiminished demand for oxygen. As these conditions persist, areas of the hypolimnion zone may become anoxic. This process is accelerated by the input of additional oxidizable material in the form of pollution or other man-made products and is reflected by an increase in the oxygen depletion rate, all other factors remaining equal. Therefore, the dissolved oxygen depletion rate is useful for assessing changes in the aquatic environment, particularly for the purpose of determining historical trends.

The mean hypolimnetic oxygen concentrations at different points in time are used to determine oxygen depletion rates. A dissolved oxygen depletion rate may also be referred to as the net oxygen demand rate per unit volume (Fay and Herdendorf, 1981). In this chapter, however, the term "dissolved oxygen depletion rate" refers to a volumetric rate calculation which requires that only the oxygen concentrations be known. Dissolved oxygen depletion rates for Lake Erie have previously been calculated in various ways. Both Charlton (1979) and Dobson and Gilbertson (1971) calculated the dissolved oxygen depletion rates in the hypolimnion zones of Lake Erie using a simple linear regression technique where the basin-wide mean hypolimnion oxygen concentration for each cruise was regressed on the time at the midpoint of that cruise. This mode of analysis has historically been the common approach to follow. The Mesolimnion Exchange Model developed by Burns (1976a) quantitatively includes the hypolimnetic waters in both basins and the oxygen dissolved in them. It also accounts for the oxygen entering or leaving a basin by horizontal and/or vertical circulation as well as corrections for different hypolimnion temperatures.

Burns and Rosa (1981) attempted to establish a representative, homogeneous area for each basin by calculating a depletion rate distribution map. Their analysis, however, used interval depletion rates. An interval depletion rate is simply the loss in oxygen (either at a station or as given by the difference in mean value of a representative area) divided by the time (in days) between two cruises. For a stratification period containing NC cruises (NC-1), interval depletion rates are averaged to obtain an overall depletion rate corresponding to the entire stratification period.

In addition to the different methods of calculating depletion rates, different criteria have been used to obtain the hypolimnetic dissolved oxygen concentration for a sampling station. Dobson and Gilbertson (1971) used the arithmetic mean of the concentration in samples for which the temperature was within 3°C of the minimum temperature. Charlton (1979) used only near-bottom values at stations that had a depth over 15 m, were stratified and showed no evidence of incursion of Eastern Basin water, but he removed data which he considered to be in error for reasons which are set out in an appendix of his paper. Both of these papers consider depletion rates for only the Central Basin. Burns and Rosa (1981) selected values on the basis of sampling depth and temperature. One point of agreement by these authors and Carr (1962) and Beeton (1963) is that some data selection is necessary due to the sampling problems and nonrepresentative sampling. The final result is conflicting conclusions about the status of hypolimnetic oxygen depletion rates over the years from 1929 to the present.

In this chapter, the problem of data selection is addressed, and appropriate methods for both estimating the depletion rates and initial concentrations and testing the hypothesis of constancy over time are presented. A clustering procedure is proposed as a semi-objective and practicable method for determining sets of observations which form a homogeneous group and which come from hypolimnetic water. The mean hypolimnetic dissolved oxygen concentrations are calculated for the homogeneous groups, thus reducing the variability due to spatial variation in concentrations and non-representative sampling. Linear regression, taking into account the precision of the means of the groups, is used to estimate depletion rates and to test constancy over time by using models which incorporate the time structure of the data.

DATA AND METHODS OF ANALYSIS

The 1967 to 1980 data collected by the Canada Centre for Inland Waters were used, and for these data the dissolved oxygen concentration was determined by the Winkler method (e.g. Wetzel and Likens, 1979). Water temperature was chosen to be used together with the dissolved oxygen concentration, since the hypolimnion is defined in terms of temperature, and only surface and bottom pairs of values were employed for each station. Furthermore, only cruises conducted during the summer stratification period were included. Thus it was expected that groups characterizing the hypolimnion of the Central and Eastern basins should be separated from groups representing non-hypolimnion water and data from the bottom waters of stations with no stratification would be assigned to the non-hypolimnion groups.

Clustering Method

A non-hierarchical nearest-centroid clustering method with a variable number of clusters was used. The clustering algorithm (Anderson, 1982) was fashioned after the well-known ISODATA algorithm of Ball and Hall (1967), as it contains many of the same steps, but the methods within the steps differ considerably. The steps of the present algorithm, which are summarized in Figure 48, include automatic seed point generation and criteria for the elimination of clusters with too few members, for the splitting of elongated clusters and for the clumping of clusters which are too close together.

Let $X = \{x_{ij}\}$ be the data matrix where x_{ij} is the observation on the jth variable for the ith unit, and $i = 1,2...n$ and $j = 1,2...m$. Thus the units, represented by the vectors $\underline{x_i} = (x_{i1}, x_{i2}...x_{im})'$ for $i = 1,2...n$, are to be assigned to the appropriate clusters. The steps of the algorithm are described below.

1. Initialization and transformation. The user specifies k = initial number of clusters, and the minimum number of members per cluster. The first step in the calculations is the optional transformation of the observations by dividing each observation by the standard deviation of the variable, with the purpose of removing the effect of different measurement units. Accordingly, x_i, is transformed to

$$x'_{ij} = x_{ij}/s_j$$

where $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_{.j})^2$ and $\bar{x}_{.j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$. In the rest of this section the observations, either transformed or untransformed, will be denoted by x_{ij}. The initial cluster centroids or seed points, z_ℓ, for $\ell = 1,2...k$, are calculated (Northouse and Fromm, 1976) as

$$z_{1j} = \bar{x}_{.j} + \mathrm{SGN}\left\{\sin\left(\frac{2\pi}{2^j}\ell - \frac{2\pi}{2^{j+1}}\right)\right\} s_j$$

for $\ell = ,2...k$ and $j = 1,2...m$, and $\bar{x}_{.j}$ and s_j the mean and standard deviation of the jth variable, transformed or untransformed.

2. Assignment and recomputation of centroids. Each unit, $x_i = (x_{i1}, x_{i2}...x_{im})'$, is assigned to the cluster for which the Euclidean distance between the unit and the cluster centroid is minimum. Once all the units have been assigned to one of the k

clusters, the cluster centroids are recomputed. Let ω_ℓ denote the set of units constituting the ℓth cluster, then

$$z_{\ell j} = \frac{1}{n_\ell} \sum_{x_{ij} \varepsilon \omega_\ell} x_{ij}$$

where n_ℓ = number of units in cluster ℓ.

3. Too-few-members criterion. If a cluster contains fewer than the user-specified minimum, the cluster is eliminated. The members of the eliminated cluster are reassigned and the cluster centroids recomputed as described in step (2) above.

4. Splitting criterion. The ratio of a measure of scatter for the unsplit cluster to the sum of the measure of scatter for two sub-clusters is compared with a user specified threshold, θ_s, to determine whether the cluster should be split (Duda and Hart, 1973). The scatter matrix for cluster ℓ is given by

$$S_T = \sum_{\underline{x}_i \varepsilon \omega_\ell} (\underline{x}_i - z_\ell)(\underline{x}_i - z_\ell)'$$

where $\underline{z}_\ell = (z_{\ell 1}, z_{\ell 2} \ldots z_{\ell k})'$

Using the scatter matrix, the principal axis is determined and three equally spaced cuts along this axis are used to divide the cluster into pairs of sub-clusters. The measure of scatter within a cluster is obtained as the sum of the within-cluster sums of squares over all variables, which is the trace of the scatter matrix, denoted as $tr(S_T)$ for the original cluster. For the pairs of sub-clusters, the sums of the within-cluster sums of squares for both clusters are added; denote this by $\mathrm{tr}(S_W)$. The pair of sub-clusters with the smallest $\mathrm{tr}(S_W)$ is used to test for sufficient reduction in the within-cluster scatter. For

$$\frac{\mathrm{tr}(S_T)}{\min \mathrm{tr}(S_W)} > \theta_s$$

the cluster is split into the pair of sub-clusters corresponding to min $\mathrm{tr}(S_W)$ and the cluster centroids are determined for the new sub-clusters. This splitting criterion is applied separately for each original cluster.

5. Lumping criterion. The lumping parameter, ρ_K, defined by Northouse and Fromm (1976), is

$$\rho_K = \frac{m}{3K} \sum_{k=1}^{K} D_k$$

where m = number of variables, K = present number of clusters and D_k = minimum distance between cluster k and the other k-1 clusters. If any $D_k < \rho_K$, the kth cluster is eliminated. The members of the eliminated cluster are reassigned and the centroids recomputed as described in step (2).

6. Test for stability. If the final clusters are the same as those obtained in the previous iteration or the maximum number of iterations has been exceeded, then the final configuration has been reached. Otherwise, steps (2) to (5) are repeated until a final configuration is achieved. Once the final clusters have been chosen, the mean and standard deviation of the dissolved oxygen concentrations are calculated for each cluster.

Application of Clustering Method

For the present purpose $\underline{x}_i = (x_{i1}, x_{i2})'$ where x_{i1} = temperature and x_{i2} = dissolved oxygen concentration for a particular station and depth, and, for example, $\underline{x}_i$ might be the surface pair and $\underline{x}_j$, the bottom pair of data for a particular station. For a given cruise, all available data for the Central and Eastern basins would form the X matrix and thus would be separated into clusters.

The clustering method has been incorporated in a FORTRAN program, and an example of the graphical information derived from a typical computer run, cruise 7022106, is given in Figure 49. The plot of the data, with the cluster number (corresponding to the cluster to which the data point was actually assigned) used as a plotting symbol is useful as a visual aid for evaluating the performance and behaviour of the cluster analysis. Clusters 1 and 4 correspond to the data exhibiting the features of the epilimnion zone (namely the warmest water temperatures and high dissolved oxygen concentrations). Cluster 2, on the other hand, contains the relatively high dissolved oxygen concentrations of the hypolimnetic Eastern Basin having the coldest water temperatures. Cluster 3 represents the hypolimnetic oxygen levels of the Central Basin. The map shows the location of the stations forming clusters 2 and 3, the clusters of data determined to be in the hypolimnion. Corresponding to these plots, the program generates a printer output which displays progressive cluster statistics, namely, the mean and standard deviation of dissolved oxygen concentrations, temperature and depth for each cluster.

The computer program is interactive and some user intervention was practised to make adjustments to meet the objectives. These were (1) separation of clusters which mixed Eastern and Central Basin stations, (2) merging of two Central Basin hypolimnion clusters and (3) isolated misclassifications (Anderson, 1982). The clustering procedure as described above was applied to the data for the years 1967 to 1970 and these results, together with depth, were used in the classification in the years 1971 to 1980. Furthermore, within any one year, only stations common to all cruises were used. This resulted in smaller numbers of data points in most clusters and exclusion of years 1971, 1974, 1976 and 1980 (Central Basin) altogether, since too few points remained. These final clusters were used to calculate the means and standard deviations of the dissolved oxygen concentrations to be used in the depletion rate calculations.

Depletion Rate Calculations and Tests

Calculation of Depletion Rates

Observations indicate that the mean dissolved oxygen concentration in the hypolimnion decreases in a linear fashion throughout the stratification. Hence, it is appropriate to consider the model

$$y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i \qquad i = 1,2 \ldots NC \qquad (4.1)$$

where NC = number of cruises in a particular zone and in the year under study,

y_i = the mean dissolved oxygen concentration (mg/L)

x_i = corresponding mid-cruise date (Julian days > April 1)

ε_i = is a random variable with mean 0

α_0, α_1 = are unknown parameters to be estimated from the data.

The parameter α_1 is the depletion rate. The usual approach for estimating α_0 and σ is to use unweighted least squares. This is equivalent to assuming that the distribution of ε_i is normal with mean 0 and variance σ^2. This approach seems appropriate if the number of stations included in the analysis is the same, not only within the year but throughout the entire period of the study, and the standard deviation of the dissolved oxygen concentration is independent of the Julian day, i.e., the time of conducting the data collection. The number of sampling stations, although constant within the years, varies from year to year. In addition, the standard deviation varies with the Julian days within each year and with the number of sampling stations in the hypolimnion zone. To demonstrate these points, a simple linear regression was performed between the standard deviation (the dependent variable) of dissolved oxygen and the corresponding number of sampling stations for data from the hypolimnion (the independent variable) of the Central Basin. The results indicated a highly significant positive ($[P < 0.01]$ with $F_{1,40} = 11.39$) association between the standard deviation, y, and the number of stations, x. The regression equation is

$$y = 0.654 + 0.030x$$

This shows that the hypolimnion is not completely homogeneous and that the number of sampling stations available is not enough for reaching a conclusion by exact analysis. Furthermore, the regression was repeated by using both the number of stations and the Julian days. The regression equation is

$$y = 1.206 + 0.033x - 0.086z$$

where z is the Julian day. The inclusion of z is found to be highly significant at ($P \leq 0.01$) with $F_{1,39} = 12.23$. Figure 50 presents the plot of the standard deviation of the dissolved oxygen of the hypolimnion of the Eastern Basin against time in Julian days. This figure shows that the standard deviation increases linearly with time. The inconstancy of variances shows that weighted regression should be used in estimating the depletion rate and for testing different hypotheses about these rates. This demonstrates that the approaches used by Dobson and Gilbertson (1971) and Charlton (1979) are not efficient from a statistical point of view. To illustrate the effect of the inequality of variances on the calculation of the depletion rates, the analysis in this chapter was performed using three different methods: (1) simple linear regression, (2) weighted linear regression with the weights as a function of the number of sampling stations, and (3) weighted linear regression with empirical weights (i.e., weights using both the number of stations and the estimated standard deviation). These three cases correspond to the following assumptions about the errors ε_i in the model: (1) ε_i are normally distributed with mean 0 and variance σ^2; (2) ε_i are normally distributed with mean 0 and variance σ^2/n_i, where n_i is the number of stations in the ith hypolimnion cluster; and (3) ε_i are normally distributed with mean 0 and variance $\sigma_i{}^2/n_i$. In the last case, $\sigma_i{}^2$ is unknown, but can be estimated by $s_i{}^2$, the variance of the dissolved oxygen concentration in the ith hypolimnion cluster. Since $s_i{}^2$ is estimated from the data, the exact probability distributions of the estimates b_0 and b_1 of α_0 and α_1 are not exactly known; if $s_i{}^2$ is assumed to be the true value of $\sigma_i{}^2$, however, then the usual inference can be made.

Detection of Time Effects in the Oxygen Concentration

The problem of determining whether the variations in the hypolimnion dissolved oxygen concentration from year to year are real or are simply the result of chance fluctuations is the same as determining whether the parameters α_0 and α_1 of model 4.1 are constant in time. Brown et al. (1975) described a number of graphical and statistical procedures for testing this problem. In the present case these require fitting model 4.1 to the data, assuming that there are no changes in α_0 and α_1 from year to year and calculating the standardized residuals, the orthogonalized residuals, the cumulative orthogonalized residuals and the squared cumulative residuals. Then these residuals are examined graphically and tested formally for time independence.

Tests for Constancy of Depletion Rate

Formal tests for constancy are performed to evaluate the evidence against the following hypotheses:

(i) H: α_0 and α_1 are constant for the entire data record,
(ii) H_1: α_0 is constant for all years, but α_1 is allowed to vary from year to year,
(iii) H_2: α_1 is the same for all years, but α_0 varies from year to year.

The first hypothesis H is concerned with testing whether the initial oxygen concentration at the beginning of the stratification season α_0 and the depletion rate α_1 are constant for the period under study, whereas H_1 tests the constancy of α_0 but allows a variable depletion rate. Finally, H_2 examines the evidence of no change in the depletion rates. All these tests are standard procedures of linear regression analysis.

RESULTS AND DISCUSSION

The mean and standard deviations of the hypolimnetic oxygen concentration by cruise for the years 1967-1979 and 1967-1980, as determined for the common stations within a year, are given for the Central Basin and the Eastern Basin in Tables 43 and 44, respectively. The number of observations used to calculate these quantities and the mid-cruise date, in Julian days with April 1 as day 0, is also given. These tables provide the data used for the depletion rate calculations.

The results of fitting model 4.1 under the three assumptions about the distribution of ε_i, which were called simple linear regression, weighted regression and regression with empirical weights, are given in Tables 45, 46 and 47. The intercept α_0 and slope α_1 are estimated by b_0 and b_1, respectively. The values of the multiple correlation coefficient R^2 and the observed significance level for the slope are also given for the case of simple linear regression. A comparison between the oxygen depletion rates as calculated in this chapter using simple linear regression, weighted regression and regression with empirical weights indicates that the differences are not substantial. This is expected on theoretical grounds, since the estimates b_0 and b_1 of α_0 and α_1 are unbiased under all the methods. However, the problem of testing for trend in the depletion rates is highly influenced by the method used in the calculation, as will be shown later.

The depletion rates based on the regression with empirical weights, as given by b_1 in Table 47, were examined for time dependency using the residuals obtained from fitting model 4.1 and also using orthogonalized residuals and squared orthogonalized residuals. Figure 51 presents the plots of the residuals against the sequential order of the observations for the Central Basin data. Figure 51a gives the plot of the standardized and the orthogonalized residuals which indicates the non-randomness of the residuals, and Figure 51b shows the plot of the cumulative residuals, which indicates a systematic departure of the residuals from non-randomness. Figure 51c shows the plot of the squared orthogonal residuals and their expected values, which suggests a large departure from non-randomness. The same applies to the Eastern Basin, as can be seen in Figure 52.

The formal analyses performed to test the hypotheses regarding constancy of the initial dissolved oxygen concentration and the depletion rate are summarized in Table 48 for the Central and Eastern basins. The results are shown for the three different assumptions about the distribution of ε_i. It is clear that the F-values differ substantially under different assumptions. For example, the F-values for testing H_1 (i.e. variable depletion rates) are 2.67, 9.11 and 3.80 when unweighted, weighted and empirically weighted regressions, respectively, are used for the Central Basin data. Hence, if unweighted regression is used, H_1 is accepted, whereas if weighted regression is used, the differences between rates are highly significant ($P < .01$), and if the empirical weights are used, then H_1 is significant at the 5% level. This shows that the results are very sensitive to the assumption of the equality of variances. As stated previously, the analysis adopted here is based on using regression with empirical weights; therefore the discussion will be limited to this case.

For the Central Basin, the last column in Table 48 shows that the F-values associated with H, H_1 and H_2 are significant at the 5% levels. This indicates that the slopes and the intercepts differ from year to year. The corresponding results for the Eastern Basin show that H and H_2 are rejected, which implies that the depletion rates can be regarded as constants independent of year; the initial oxygen concentration at the beginning of the stratification season, however,is not constant from year to year.

These tests were of hypotheses of constancy over time against the general alternative of inconstancy. The conclusion that neither initial dissolved oxygen concentration nor depletion rate are constant in the Central Basin and that initial dissolved oxygen concentration is not constant in the Eastern Basin leads to the next step, namely the determination of the form of the time dependency. An initial step is given in Figure 53, where the depletion rates are plotted against year for the two basins and a fitted curve is also shown on the plot for the Central Basin and the average depletion rate for the Eastern Basin. The quadratic polynomial fitted to the Central Basin depletion rates summarizes the tendency to higher rates at both ends of the time period.

It is of interest to discover which factors may have influenced dissolved oxygen depletion rates in the Central Basin. Factors that may have contributed to the low oxygen depletion rates in the years 1972, 1973, 1975 and 1978 are (1) high water levels in 1972, 1973 and 1975, (2) large hypolimnion thickness in 1975 and 1978; (3) low hypolimnion water temperatures in 1975; and (4) cold hypolimnion water temperatures at the onset of the summer stratification period in 1975 and 1978. The large depletion rate in 1977 may have been influenced by the warm hypolimnion temperature at the beginning of the stratification period. The incorporation of terms for some of these factors in a model for the oxygen concentration in the Central Basin is considered by El-Shaarawi in Chapter 5.

Carr (1962) suggested that oxygen consumption was increasing in the hypolimnion of the Central Basin. Dobson and Gilbertson (1971) agreed with this general conclusion and calculated a long-term trend of 0.075 mg/L/month/year for the years 1930-1970. They attributed the trend to the quantity of phytoplankton sedimented to the lake bottom. Charlton (1979), on the other hand, concluded that there was no significant trend in the dissolved oxygen depletion rate by standardizing this parameter to account for physical factors, namely water temperature and hypolimnion thickness. Burns and Rosa (1981) supported the hypothesis of a trend by accounting for other physical parameters including temperature, vertical mixing, and incoming oxygen from the Eastern Basin. This chapter does not add another conclusion to this question; it discusses the steps that precede this final conclusion.

A cluster analysis is presented which automatically separates the data into the various naturally occurring thermal and spatial regimes. The resulting dissolved oxygen depletion rate, calculated on the basis of common stations in these clusters within a year, is compared with the values obtained by others in Figure 54. The values calculated in this chapter are generally between those calculated by other authors and, with one exception, below those calculated by Fay and Herdendorf. Thus, the cluster analysis fulfills the purpose of providing a semi-objective and practicable alternative to the previous methods of data selection. Several other features of the clustering algorithm are the following. The high spatial variability is reduced, enabling the temporal effects of interest to be more significant. In general, the clustering algorithm performed well, with the poorest results being achieved with the data corresponding to the beginning of the summer stratification period, and the performance improving as the thermocline became more pronounced.

The importance of the methods of regression analysis used here to estimate the initial dissolved oxygen concentration and the depletion rate and to test hypotheses about their constancy is two-fold. First, the regression analysis should be the weighted analysis to reflect the features of the data collection properly. Secondly, the model describing depletion rate changes can be formulated so that estimation of the depletion rate is made from a model which incorporates features such as time structure.

Table 43. **Summary of Dissolved Oxygen Concentrations and Depths for the Groups Restricted to Common Stations within a Year, Central Basin 1967 to 1979**

Year	Cruise	No. of points	Julian days (> April 1)	Dissolved oxygen (mg/L)		Depth (m)	
				Mean	S.D.	Mean	S.D.
1967	6722103	3	83.0	8.55	0.51	20.00	1.73
	6722104	3	98.0	7.49	1.09	21.00	1.73
	6722105	3	105.0	5.88	0.60	22.00	0.00
	6722107	3	125.0	4.59	0.97	20.00	1.73
	6722109	3	146.0	1.70	0.77	19.00	0.00
1968	6822102	9	49.0	11.77	0.41	19.56	3.20
	6822104	9	77.5	8.45	0.36	20.11	3.21
	6822108	9	122.0	4.24	1.22	19.67	3.27
1969	6922103	11	62.0	10.02	0.58	20.45	2.06
	6922104	11	94.0	6.60	0.56	20.81	1.53
	6922105	11	121.5	3.36	1.94	20.73	1.61
	6922107	11	147.5	0.97	0.59	21.18	2.08
1970	7022103	9	37.0	13.56	0.29	19.33	1.73
	7022104	9	64.0	10.18	0.52	20.88	1.05
	7022106	9	95.5	6.45	0.36	20.55	0.88
	7022109	9	148.0	1.39	0.95	21.66	1.41
1972	7222101	12	25.5	13.25	0.38	19.58	2.07
	7222102	12	68.0	8.60	2.37	20.42	1.49
	7222103	12	88.5	9.01	0.43	20.33	1.50
	7222104	12	124.5	4.50	1.14	20.25	1.65
	7222106	12	151.0	1.57	0.68	20.08	2.15
1973	7322101	28	13.0	13.36	0.39	18.80	2.23
	7322103	28	117.0	4.39	1.54	19.00	2.55
	7322106	20	150.5	2.46	2.24	19.41	2.46
1975	7522102	32	46.0	12.52	0.80	19.97	2.02
	7522106	32	87.0	9.30	1.53	20.01	2.06
	7522107	32	129.0	5.80	2.32	19.92	2.26
1977	7722102	35	63.0	10.30	0.89	20.41	2.14
	7722103	35	75.5	8.85	1.33	20.23	2.08
	7722105	35	96.5	7.14	1.34	20.28	2.06
	7722106	35	117.5	4.88	1.24	20.07	2.09
	7722107	35	131.5	3.62	1.91	20.11	2.11
1978	7822103	28	60.5	11.14	1.16	19.75	2.40
	7822104	28	82.5	9.45	1.91	19.99	2.31
	7822106	28	106.0	7.71	2.80	19.75	2.67
	7822108	28	124.0	6.40	2.69	19.61	2.42
	7822110	28	143.0	4.69	1.98	19.76	2.47
1979	7922103	9	46.0	12.66	0.50	21.00	1.54
	7922104	9	73.0	9.66	0.57	21.22	1.35
	7922106	9	95.0	7.07	1.16	20.94	1.38
	7922109	9	115.0	4.58	1.85	21.34	1.30
	7922112	9	144.5	2.48	1.73	21.44	1.16

Table 44. Summary of Dissolved Oxygen Concentrations and Depths for the Groups Restricted to Common Stations within a Year, Eastern Basin 1967 to 1980

Year	Cruise	Number of points	Julian days (> April 1)	Dissolved Oxygen (mg/L)		Depth (m)	
				Mean	S.D.	Mean	S.D.
1967	6722101	7	61.5	12.05	0.529	38.21	11.73
	6722103	7	87.5	11.44	0.35	38.57	11.28
	6722105	7	102.0	10.19	0.35	30.43	10.92
	6722107	7	122.5	9.69	0.73	38.29	13.29
	6722109	7	143.5	8.57	1.05	34.57	11.53
	6722111	7	164.5	7.78	1.04	38.71	10.79
1968	6822102	4	47.0	12.70	0.23	44.00	5.94
	6822104	4	76.0	11.52	0.09	45.76	2.50
	6822108	4	120.0	9.35	0.24	49.00	8.12
	6822111	4	181.0	6.62	0.41	47.75	8.22
1969	6922103	3	60.0	12.88	0.35	52.00	10.39
	6922104	3	96.0	11.11	0.44	51.00	10.44
	6922105	3	122.0	9.87	0.16	51.00	8.66
	6922107	3	150.0	9.21	0.21	52.00	9.54
	6922108	3	166.0	7.98	0.72	49.68	8.96
	6922110	3	197.0	6.85	0.43	51.67	8.62
1970	7022103	3	35.0	13.64	0.06	39.00	9.54
	7022104	3	62.0	12.61	0.14	38.67	7.50
	7022106	3	93.0	11.07	0.83	41.00	9.54
	7022107	3	119.0	9.60	1.24	40.00	8.72
	7022109	3	147.0	9.27	0.22	41.67	8.39
	7022111	3	175.0	7.29	0.57	41.33	8.15
1972	7222101	5	24.0	13.78	0.45	35.20	13.16
	7222102	5	67.0	12.75	0.88	35.40	13.26
	7222103	5	87.0	11.53	0.75	35.20	10.57
	7222104	5	123.0	9.06	0.73	35.40	12.04
	7222107	5	181.0	6.95	1.36	36.40	11.85
1973	7322101	4	11.0	13.53	0.20	38.25	14.57
	7322103	4	115.5	9.51	0.88	38.25	13.89
	7322106	4	149.5	8.07	0.71	38.75	11.70
1975	7522102	9	46.0	13.43	0.73	44.72	9.51
	7522106	9	86.5	11.41	0.56	43.61	7.37
	7522107	9	129.0	9.83	0.60	42.56	6.21
	7522110	9	191.0	8.29	0.76	42.22	7.24
	7522111	9	211.0	6.72	0.72	43.47	8.13
1977	7722102	12	60.5	12.55	0.28	40.58	6.11
	7722103	13	74.0	11.81	0.21	41.15	6.50
	7722105	13	95.0	10.76	0.78	41.46	5.68
	7722106	13	116.0	9.77	0.28	40.38	5.11
	7722107	13	130.0	9.08	0.51	41.15	5.58
	7722109	13	151.0	8.57	0.61	39.92	5.14
	7722111	13	172.0	5.67	0.69	36.69	5.27
1978	7822103	10	59.0	13.23	0.50	39.70	4.72
	7822104	13	80.5	11.83	0.50	40.69	5.33
	7822106	13	104.5	11.29	0.40	41.19	4.82
	7822108	13	122.0	10.63	0.12	40.07	4.44
	7822110	13	141.5	9.75	0.65	38.77	4.92
	7822111	13	169.0	9.03	0.54	41.35	4.72
	7822114	13	183.0	8.30	0.45	39.31	5.02
1979	7922103	4	45.0	13.44	0.14	41.50	2.89
	7922104	3	71.0	12.64	0.19	41.33	3.51
	7922106	4	93.5	11.92	0.26	41.75	2.36
	7922109	4	113.5	10.73	0.43	41.75	2.87
	7922112	4	145.0	9.73	0.26	44.00	4.89
	7922114	4	178.0	8.50	0.34	46.75	9.74
1980	8022101	2	98.0	11.20		42.00	4.24
	8022102	2	107.0	10.75	0.04	47.50	3.54
	8022104	2	133.0	9.54	0.03	40.50	2.12
	8022104	2	142.0	10.12	0.76	41.50	4.95
	8022105	2	155.0	9.00	0.45	43.00	4.24

Table 45. Summary of Simple Linear Regression Parameters and Precision Indicators from Regression Analysis of Dissolved Oxygen Concentrations, Central and Eastern Basins

	Year	Estimated slope* b_1	Estimated intercept b_0	Multiple correlation coefficient R^2	Significance level for slope α
Central	1967	0.1078	17.6529	0.9790	0.005
	1968	0.1023	16.6277	0.9966	0.05
	1969	0.1071	16.6183	0.9981	0.005
	1970	0.1092	17.3028	0.9957	0.005
	1972	0.0903	15.6452	0.9644	0.005
	1973	0.0809	14.3037	0.9952	0.05
	1975	0.0810	16.2783	0.9997	0.025
	1977	0.0967	16.3164	0.9979	0.005
	1978	0.0772	15.8400	0.9986	0.005
	1979	0.1059	17.3228	0.9905	0.005
Eastern	1967	0.0430	14.8409	0.9819	0.005
	1968	0.0459	14.9075	0.9993	0.005
	1969	0.0435	15.3807	0.9905	0.005
	1970	0.0441	15.2171	0.9821	0.005
	1972	0.0464	15.2841	0.9732	0.005
	1973	0.0392	13.9762	0.9996	0.025
	1975	0.0374	14.9026	0.9797	0.005
	1977	0.0553	16.0469	0.9486	0.005
	1978	0.0373	15.1644	0.9867	0.005
	1979	0.0381	15.2578	0.9936	0.005
	1980	0.0345	14.4974	0.8602	0.025

* All slopes are negative.

Table 46. The Estimates of the Slopes and Intercepts Using Weighted Regression

	Central Basin		Eastern Basin	
Year	Slope*	Intercept	Slope*	Intercept
1967	0.1077	17.645	0.0430	14.838
1968	0.1024	16.636	0.0458	14.905
1969	0.1071	16.618	0.0425	15.240
1970	0.1092	17.303	0.0441	15.217
1972	0.0903	15.645	0.0464	15.284
1973	0.0814	14.321	0.0392	13.976
1975	0.0810	16.278	0.0374	14.903
1977	0.0967	16.317	0.0553	16.056
1978	0.0772	15.840	0.0370	15.110
1979	0.1059	17.323	0.0380	15.246
1980	—	—	0.0345	14.498

* All slopes are negative.

Table 47. The Estimates of the Slopes and Intercepts Using Empirical Weighted Regression

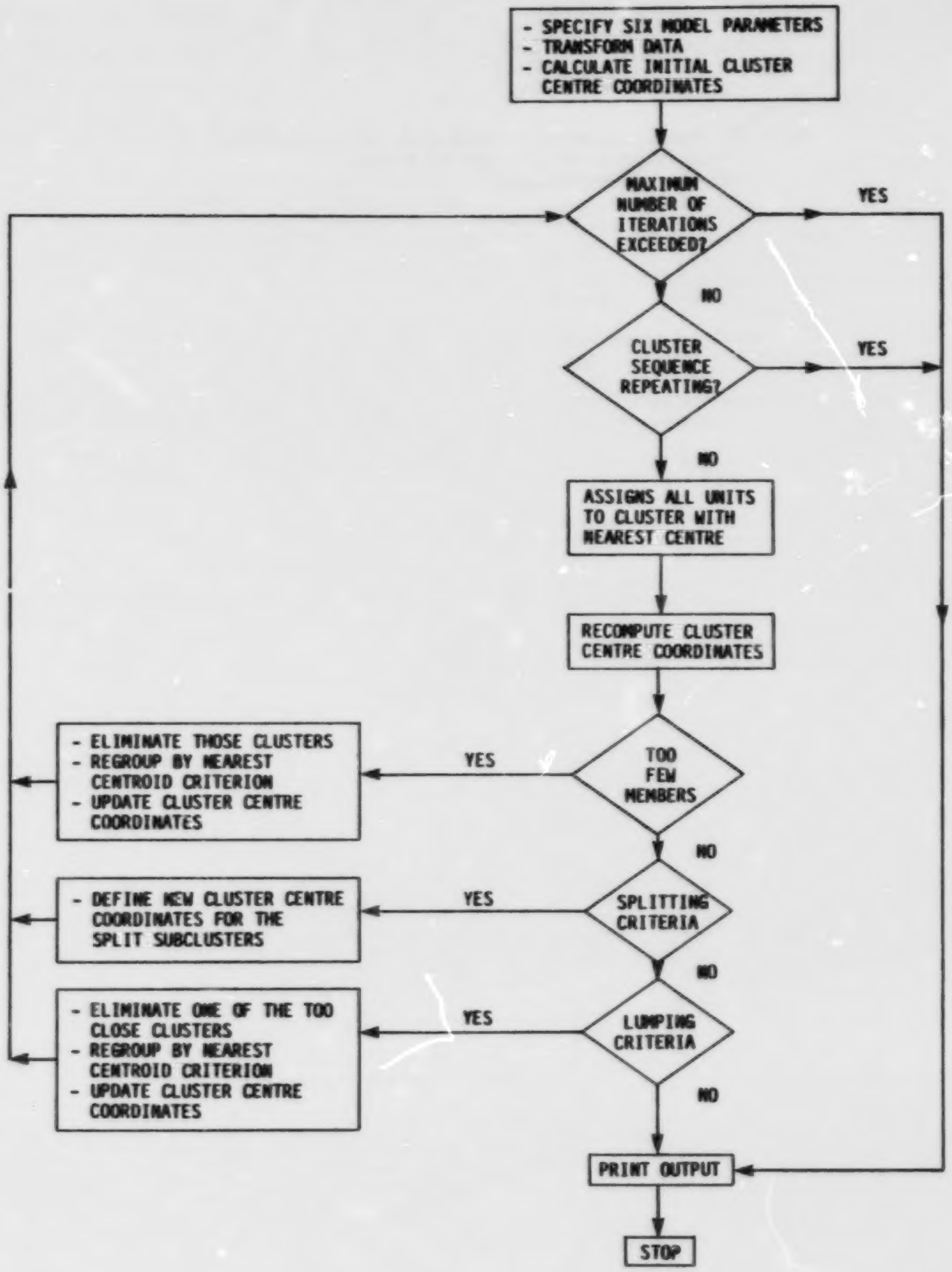
	Central Basin		Eastern Basin	
Year	Slope*	Intercept	Slope*	Intercept
1967	0.1070	17.415	0.0444	14.993
1968	0.1089	17.023	0.0464	15.011
1969	0.1060	16.572	0.0410	15.072
1970	0.1165	17.775	0.0402	15.057
1972	0.0875	15.817	0.0446	15.010
1973	0.0836	14.442	0.0392	13.962
1975	0.0802	16.219	0.0373	14.824
1977	0.0975	16.401	0.0520	15.705
1978	0.0773	15.828	0.0371	15.148
1979	0.1090	17.415	0.0372	15.175
1980	—	—	0.0469	15.787

* All slopes are negative.

Table 48. Results of Testing the Equality of the Slopes and the Equality of the Intercepts for Model 4.1

Basin	Model	Degrees of freedom	Unweighted F-value	Weighted F-value	Empirical weight F-value
Central	Constant slope and different intercepts	9,22	3.62*	4.05*	4.47*
	Different slopes and constant intercept	9,22	2.67	9.11**	3.80*
	Constant slope and constant intercept	18,22	6.58**	7.77**	11.97**
Eastern	Constant slope and different intercepts	10,38	2.59*	3.25*	4.63**
	Different slopes and constant intercept	10,38	1.55	1.30	1.38
	Constant slope and constant intercept	20,38	6.59**	6.74**	15.15**

* Indicates significance at the 5% level.
** Indicates significance at the 1% level.



- SPECIFY SIX MODEL PARAMETERS
- TRANSFORM DATA
- CALCULATE INITIAL CLUSTER CENTRE COORDINATES
MAXIMUM NUMBER OF ITERATIONS EXCEEDED?
YES
NO
CLUSTER SEQUENCE REPEATING?
YES
NO
ASSIGNS ALL UNITS TO CLUSTER WITH NEAREST CENTRE
RECOMPUTE CLUSTER CENTRE COORDINATES
TOO FEW MEMBERS
YES
- ELIMINATE THOSE CLUSTERS
- REGROUP BY NEAREST CENTROID CRITERION
- UPDATE CLUSTER CENTRE COORDINATES
NO
SPLITTING CRITERIA
YES
- DEFINE NEW CLUSTER CENTRE COORDINATES FOR THE SPLIT SUBCLUSTERS
NO
LUMPING CRITERIA
YES
- ELIMINATE ONE OF THE TOO CLOSE CLUSTERS
- REGROUP BY NEAREST CENTROID CRITERION
- UPDATE CLUSTER CENTRE COORDINATES
NO
PRINT OUTPUT
STOP

Figure 48. Flow diagram of the clustering algorithm.

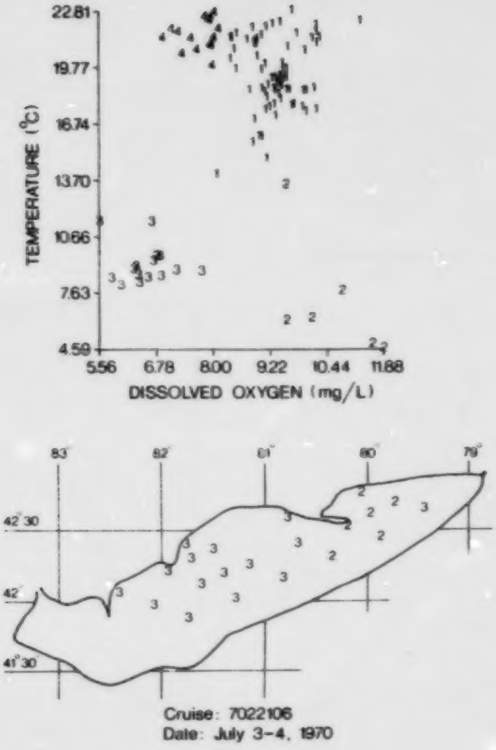


Figure 49. Example of computer plots produced by the clustering program for cruise 7022106.

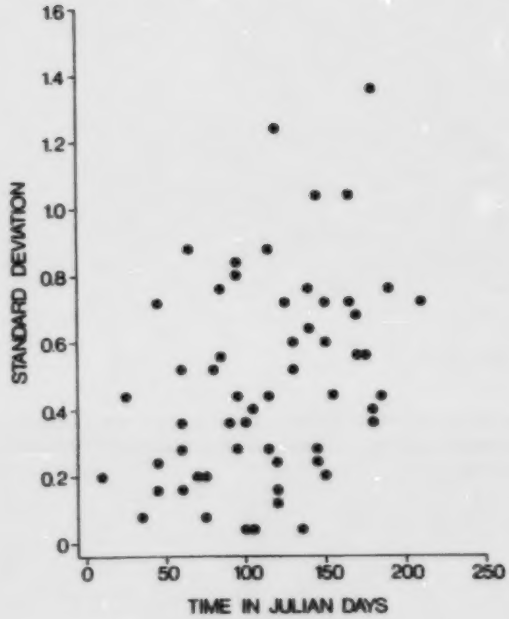


Figure 50. Standard deviation of the hypolimnetic oxygen concentration for clusters in the Eastern Basin plotted against Julian days.

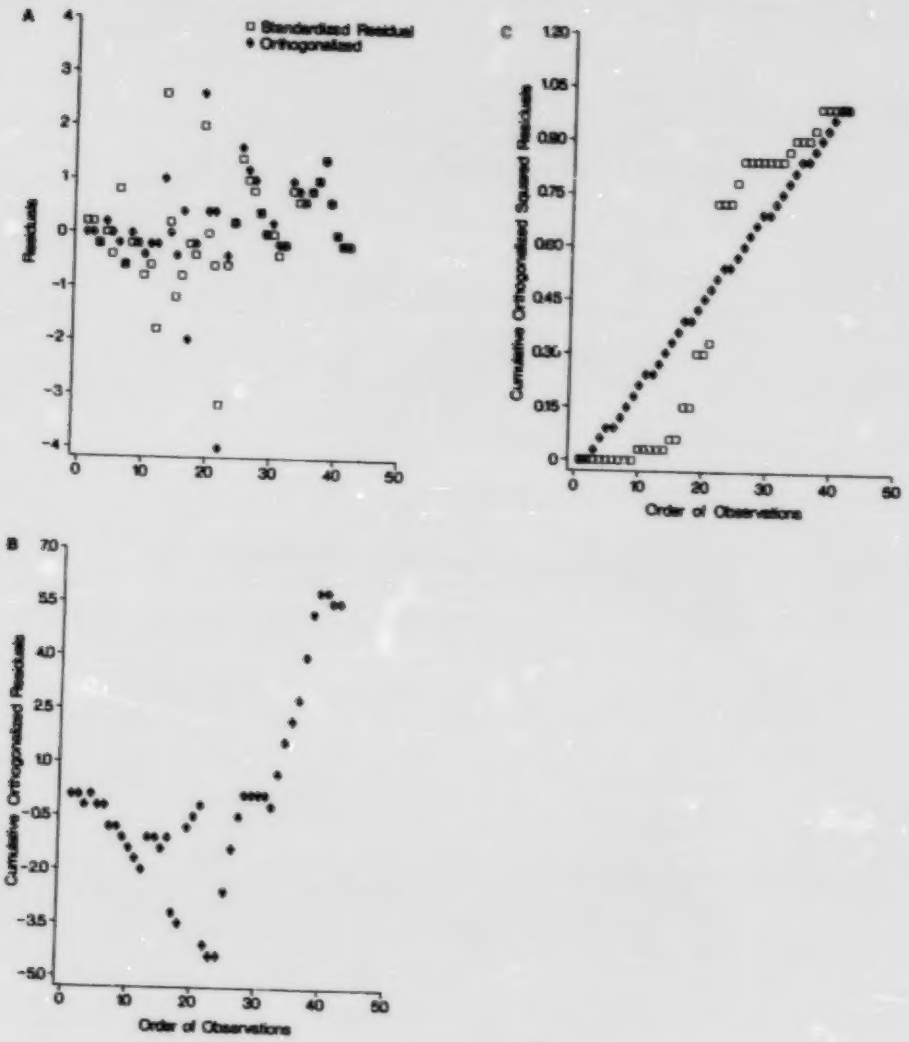


Figure 51. Plots of residuals against sequential order of the observations for the Central Basin: *a* — standardized and orthogonalized residuals; *b* — cumulative residuals; *c* — squared orthogonalized residuals.

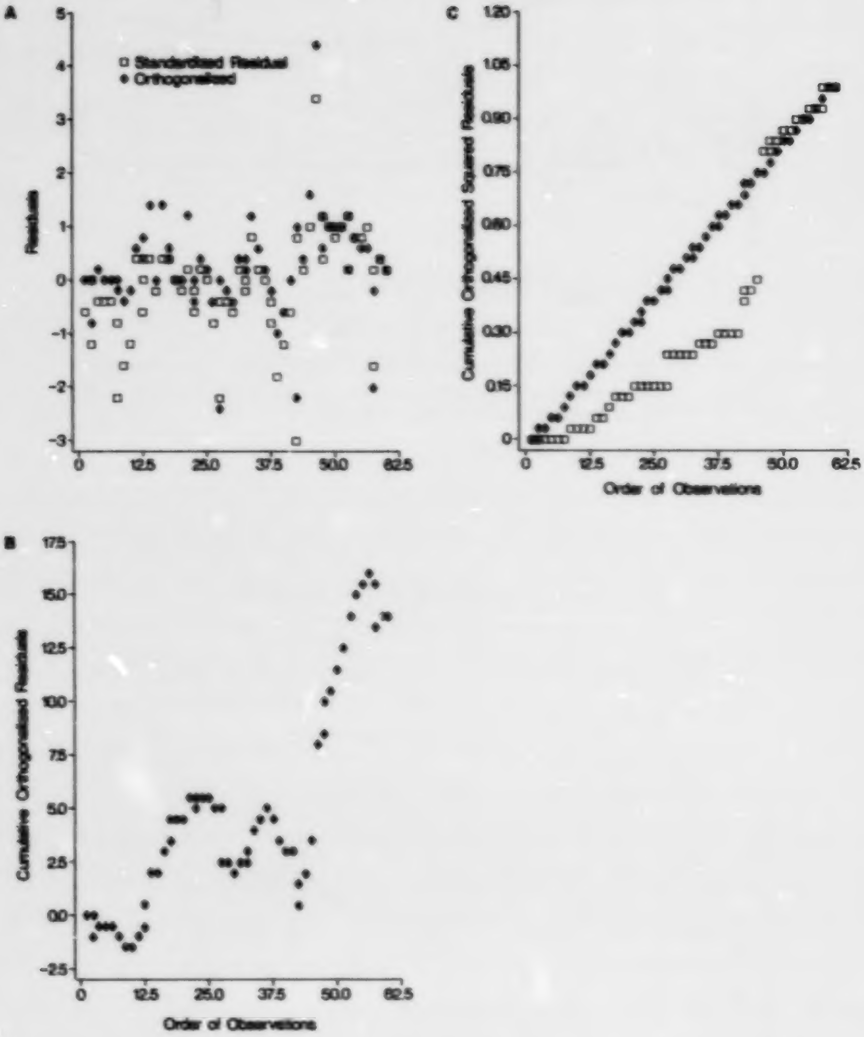


Figure 52. Plots of residuals against sequential order of the observations for the Eastern Basin: *a* — standardized and orthogonalized residuals; *b* — cumulative residuals; *c* — squared orthogonalized residuals.

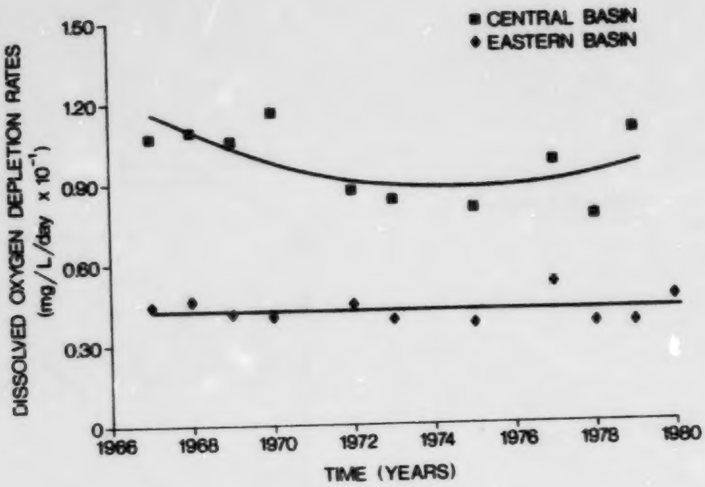


Figure 53. Dissolved oxygen depletion rates against year with fitted curves for the Central and Eastern basins.

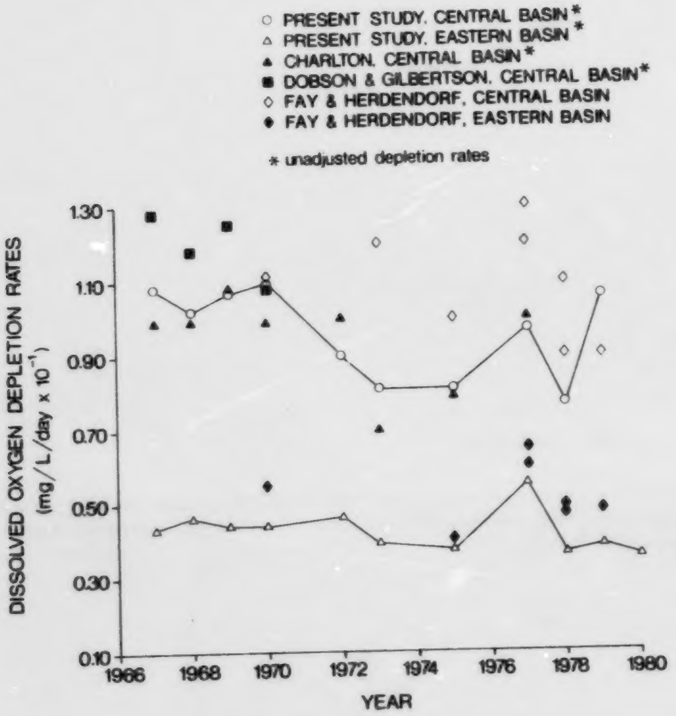


Figure 54. Comparison of uncorrected dissolved oxygen depletion rates calculated by several authors.

CHAPTER 5

A Statistical Model for Dissolved Oxygen in the Central Basin of Lake Erie

by A.H. El-Shaarawi

INTRODUCTION

Anderson et al. (Chapter 4) demonstrated that at any specified time within the summer stratification period and during the years 1967 to 1979, there were statistically significant differences from year to year in the oxygen concentration of the hypolimnion of the Central Basin of Lake Erie. Since the mean oxygen concentration of a hypolimnetic basin decreases in a linear fashion throughout the stratification, these year to year differences can be separated into two correlated but distinct components. The first of these components is due to the variations of the intercept of the line, while the second is due to the variations of the slope of the line. It is more appropriate to refer to the intercept and the slope as the initial oxygen concentration at the onset of the stratification period and the depletion rate, respectively. In Chapter 4, no attempt was made to explain the variations in the initial oxygen concentration or in the depletion rate, but the authors summarized the pattern of the trend in the depletion rate by a quadratic equation. Dobson and Gilbertson (1971) concluded that the 1970 depletion rate was more than double the rate estimated for 1929. They attributed the increase in the depletion rate to increases in phytoplankton production caused by increased nutrient inputs. Charlton (1980a,b) concluded that there was no long-term trend in the depletion rate and the apparent differences between the rates were mostly related to variations in the hypolimnetic thickness and temperature. Barica (1982) stated that Charlton's results had recently been challenged by F. Rosa and N.M. Burns. It should be mentioned that with the exception of Anderson et al. (Chapter 4), none of the above-mentioned authors has performed or provided a statistical test for trend in the depletion rates.

Thienemann (1915) distinguished between the effects of five factors on the depletion rate: (1) the season of the year, (2) the position of the lake (with respect to prevailing winds), (3) the magnitude of the volume constituting the hypolimnion, and the ratio of the water above and below the thermocline, (4) the temperature of the hypolimnetic waters, and (5) the quantity of organic matter transported into the bottom waters. Clearly, the maximum that man can do is to control the effect of the last factor, especially by controlling industrial and agricultural pollution. It is recognized that phosphorus limits the aquatic plant productivity in the Great Lakes. To control pollution of the Great Lakes, the International Joint Commission developed phosphorus loading objectives for each of the Great Lakes. Slater and Bangay (1980) stated that in Lake Erie, the point source load had been reduced from approximately 10 000 metric tons/yr in 1972/73 to 5700 tons/yr in 1977, and it is expected to be 2100 metric tons/yr when all controls are in place to meet the phosphorus effluent limit of 1 mg/L. It is important to study the effect of these remedial steps on dissolved oxygen concentration in the hypolimnion of Lake Erie.

In this chapter, a statistical model is developed for dissolved oxygen concentrations in the hypolimnion. The model uses the lake water level, the hypolimnion temperature and the average total phosphorus concentrations as explanatory variables. The probability of the occurrence of anoxia in the Central Basin is estimated from the model as a function of phosphorus, temperature and water level. Furthermore, the model is used to predict the phosphorus level for a given depletion rate and water level. The use of the model for regulation and for setting standards is also given.

MODEL DEVELOPMENT

The results of Anderson et al. (Chapter 4) demonstrate that the initial hypolimnetic dissolved oxygen concentration α_0 and the depletion rate α_1 vary from year to year. As indicated previously, there are many factors causing this variation. The influences of the water level, the temperature of the hypolimnetic water and the yearly mean total phosphorus concentration are examined here and used to model the dissolved oxygen concentrations in the Central Basin of Lake Erie.

Figures 55 and 56 give the plots of the estimated intercept b_0 and the estimated depletion rate b_1 against the Lake Erie water level. The values of b_0 and b_1 represent the estimates of the initial dissolved oxygen concentration at the beginning of the stratification period and the oxygen depletion rates and are the estimates of (the parameter) α_0 and α_1 in model 5.1 of Anderson et al. for the Central and Eastern basins. This model is given as

$$y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i \qquad i = 1,2...NC \qquad (5.1)$$

where NC is the number of cruises in the year under study, y_i is the mean O_2 concentration (mg/L), x_i is the mid-cruise date (Julian days > April) and ε_i is a random variable with mean 0.

The Eastern Basin values do not indicate the presence of an association between the water level and either b_0 or b_1; the values of b_0 and b_1 of the Central Basin, however, show a strong negative association with the water level. Since the values of b_0 appear to be approximately linearly associated with the water level, it was decided to determine whether the fluctuations of the intercepts b_0 could be modelled using the water level as a covariate. Hence, model 5.1 is modified to

$$y_{ij} = \beta_0 + \beta_1 w_j + \gamma_j x_{ij} + \varepsilon_{ij} \qquad \begin{array}{l} i = 1,2\ldots NC_j \\ j = 1,2\ldots m \end{array} \qquad (5.2)$$

where NC_j is the number of cruises in the jth year; m is the number of years; y_{ij} is the mean dissolved oxygen concentration in the ith cruise during the jth year; x_{ij} is the mid-cruise date in Julian days for the ith cruise in the jth year; w_j is the water level in the jth year; ε_{ij} is a normal random variable with mean 0 and estimated variances S_{ij}/N_{ij}, where S_{ij} is the standard deviation of the dissolved oxygen during the jth year and N_{ij} is the corresponding number of sampling stations; and $\beta_0, \beta_1, \gamma_1 \ldots \gamma_m$ are the unknown parameters of the model. The parameters $\gamma_1 \ldots \gamma_m$ are the adjusted oxygen depletion rates where m is equal to 10 in the present study.

Model 5.2 was fitted to the Central Basin data and the results indicate that the inclusion of the water level explains entirely the year to year variability in the intercepts of model 5.2. The statistic

$$LF = 22\{RES(2) - RES(1)\}/\{8 \times RES(1)\}$$

measures the lack of fit of model 5.2 as compared with model 5.1, where RES (1) and RES (2) are the residual sums of squares under models 5.1 and 5.2, respectively. The statistic LF has an F-distribution with 8 and 22 degrees of freedom. The observed LF value is 0.95, which is not significant. Table 49 gives the estimates of the parameters of model 5.2. The quantities $\hat{\gamma}_1, \hat{\gamma}_2 \ldots \hat{\gamma}_{10}$ represent the estimate values of the dissolved oxygen depletion rates $\gamma_1, \gamma_2 \ldots \gamma_{10}$ for the ten years under study. From the table it appears that $\hat{\gamma}_1, \hat{\gamma}_2 \ldots \hat{\gamma}_{10}$ vary substantially from year to year. The equality of the γ_i's was tested using an F-statistic, and the results indicate that the differences between the depletion rates are highly significant ($P < .01$ and $F_{9,30} = 12.475$). The estimated values $\hat{\gamma}_1 \ldots \hat{\gamma}_{10}$ are plotted against the corresponding water level in Figure 57. The plot shows a strong (approximately linear) negative association between them. Hence, the values of γ_j are modelled as:

$$\gamma_j = \Theta_0 + \Theta_1 (w_j - \bar{w}) \tag{5.3}$$

where Θ_0 and Θ_1 are constant, w_j is the lake water level in feet during the jth year, and $\bar{w}$ is the mean of the yearly water levels during the study period. The values of γ_j in model 5.2 are replaced by the right-hand side of expression 5.3. This leads to the new model

$$\begin{aligned} y_{ij} = {} & \beta_0 + \beta_1 (w_j - \bar{w}) + \Theta_0 (x_{ij} - \bar{x}) + \Theta_1 (w_j - \bar{w}) \\ & (x_{ij} - \bar{x}) + \varepsilon_{ij}, \qquad i = 1,2 \ldots NC_j \\ & \qquad\qquad\qquad\quad j = 1,2 \ldots m \end{aligned} \tag{5.4}$$

where $\bar{x}$ is the mean of the cruise Julian date for all the years. The estimated regression equation is:

$$y_{ij} = 7.0043 + .4147\ (w_j - \bar{w}) - .1007\ (x_{ij} - \bar{x})$$
$$+ .0233\ (x_{ij} - \bar{x})(w_j - \bar{w})$$

The lack of fit test for comparing model 5.2 with model 5.4 indicated that model 5.4 is not adequate ($P < .01$, $F_{8,30} = 6.80$), since it does not completely explain the variation in the depletion rates.

Further attempts to account for the remaining variabilities in the depletion rates led to the use of the temperature of the hypolimnion. It was noted in El-Shaarawi (1984a) that hypolimnion temperature fluctuates from year to year and sometimes varies within the year. In fact, the temperature values were represented by a straight line when regressed against time; hence, a more representative value for the mean hypolimnion temperature is obtained from the regression line when the Julian day is set at day 200 (note that the stratification period is taken to be between Julian days 150 and 250). If the hypolimnion temperature on day 200 in the jth year is denoted by T_j and the mean temperature over all years by $\bar{T}$, then it is found that model 5.4 should be modified to:

$$y_{ij} = \beta_0 + \beta_1\ (w_j - \bar{w}) + \Theta_1\ (x_{ij} - \bar{x}) + \Theta_1\ (w_j - \bar{w}) \qquad (5.5)$$
$$(x_{ij} - \bar{x}) + \alpha_1\ (T_j - \bar{T}) + \varepsilon_{ij}$$

The estimated value of the dissolved oxygen concentration is:

$$\hat{y}_{ij} = 7.0921 + .1522\,(w_j - \bar{w}) - .10085\,(x_{ij} - \bar{x}) + .020881\,(x_{ij} - \bar{x})\,(w_j - \bar{w}) - .28821\,(T_j - \bar{T}) \tag{5.6}$$

The lack of fit statistic is $F = 4.64$, which is found to be significant when compared with the F-distribution with 7 and 30 degrees of freedom. Additional plots and an examination of the residuals revealed that model 5.6 should include an interaction term between the hypolimnion temperature and the water level. This resulted in the estimated form of the model for the dissolved oxygen as:

$$\hat{y}_{ij} = 7.011 + .1128\,(w_j - \bar{w}) - .10015\,(x_{ij} - \bar{x}) + .017287\,(x_{ij} - \bar{x})(w_j - \bar{w}) - .407859\,(T_j - \bar{T}) - .53079\,(T_j - \bar{T})\,(w_j - \bar{w}) \tag{5.7}$$

The lack of fit statistic is $F = 2.704$, which is significant at the 5% level when compared with the F-distribution with 6 and 30 degrees of freedom. It was decided to check whether the presence of the interaction term between the hypolimnion temperature and the water level in model 5.7 reduces the importance of the main effect of the water level (the term with the water level only) in the model. Performing the required F-test resulted in the acceptance of the hypothesis that the water level term is not needed in model 5.7. The observed F-value is 0.3613, which is not significant at the 5% level. As a result, the estimated dissolved oxygen concentration as a function of the lake water level and the hypolimnetic temperature is

$$\hat{y}_{ij} = 7.014 - .10005\,(x_{ij} - \bar{x}) + .015896\,(w_j - \bar{w})(x_{ij} - \bar{x}) \qquad (5.8)$$
$$- .42064(T_j - \bar{T}) - .535174\,(T_j - \bar{T})(w_j - \bar{w})$$

The lack of fit statistic associated with model T is 2.58, which is significant at the 5% level.

This analysis indicates that the water level and the hypolimnetic temperature are not completely capable of explaining all the non-random variabilities in the hypolimnetic oxygen concentration. Hence, an attempt is made in the remaining part of this section to determine the effect of the process of eutrophication on the oxygen concentration. Let P_j $(j = 1,2...10)$ be the mean total phosphorus concentration (TP) of the epilimnion of the Central Basin of Lake Erie during the jth year. The inclusion of TP in the model led to the modification of model 5.8 to

$$\hat{y}_{ij} = 7.063 - .10073\,(x_{ij} - \bar{x}) + .018059\,(w_j - \bar{w})(x_{ij} - \bar{x}) \qquad (5.9)$$
$$- .3436(T_j - \bar{T}) - .4677\,(T_j - \bar{T})(w_j - \bar{w}) - 51.1339(P_j - \bar{P})$$

where $\bar{P}$ is the arithmetic mean of $P_1 \ldots P_{10}$. The lack of fit statistic associated with model 5.9 is 1.84, which is not significant at the 5%

level when compared with the F-distribution with 6 and 30 degrees of freedom. Although model 5.9 explains most of the significant variability in the data, it was felt that TP should not only influence the hypolimnetic initial oxygen concentration at the beginning of the stratification season but should also influence the oxygen depletion rate. The residual sums of squares under model 5.2 and model 5.9 are 165.572 and 226.428, respectively, which also shows that the magnitude of the difference between the residual sum of squares of the two models, although not statistically significant, is large. Hence in model 5.9 an additional term was added to represent the interaction between time in Julian days and TP. This resulted in a substantial reduction of the residual sum of squares from 226.428 to 194.654. The new model is given by

$$\begin{aligned}\hat{y}_{ij} = {} & 7.1037 - .09778\,(x_{ij} - \bar{x}) + .01448\,(x_{ij} - \bar{x})(w_j - \bar{w}) \\ & - .36668(T_j - \bar{T}) - .4978(T_j - \bar{T})(w_j - \bar{w}) - 71.4513(P_j - \bar{P}) \\ & - 1.4575\,(x_{ij} - \bar{x})(P_j - \bar{P})\end{aligned} \tag{5.10}$$

and the associated lack of fit statistic is 1.054, which is not significant and is close to one (the theoretical expectation of the lack of fit statistic). Hence, model 5.10 is taken as the final model for the oxygen concentration of the hypolimnion of the Central Basin. The values of $\bar{x}$, $\bar{T}$, $\bar{w}$, and $\bar{p}$ are 96.393, 10.393, 571.881 and .0180, respectively. Assuming that model 5.10 is the true model, the dependence of the depletion rate on TP is significant at the 5% level ($t_{35} = 2.391$), and the dependence of oxygen concentration on the interaction between the temperature and the water level is highly significant ($t_{35} = 3.52$). Furthermore, the effect of TP on the oxygen concentration is extremely significant ($F_{2,35} = 6.333$). The inclusion of TP in the model resulted in changing the residual sum of squares from 265.09 to 194.654, which is a very large reduction. Figure 58 shows the plot of the observed oxygen concentration and the estimated regression line against time in Julian days when April 1 is taken as day 0. The plots show that the model fits the data reasonably well.

It should be emphasized that model 5.10 shows that during the period 1967 to 1979 and using data collected only by CCIW, the fluctuations in oxygen concentration can be explained by the water level, hypolimnetic temperature and TP. In the rest of this chapter, the model is used to (i) predict the depletion rate as a function of the water level and TP; (ii) describe the trend in historical data; (iii) estimate the level of TP from the depletion rate, the water level and the hypolimnetic temperature; and (iv) estimate the probability of anoxia for different water levels, hypolimnion temperatures, total phosphorus and length of stratification period.

PREDICTION OF THE DEPLETION RATE AND TREND IN THE HISTORICAL RECORD

Examination of model 5.10 reveals that the negative sign of the coefficient of $(x_{ij} - \bar{x})$ represents the estimate of the depletion rate in the jth year, and hence is given by

$$D_j = 0.09778 - .01448\ (w_j - \bar{w}) + 1.4575\ (P_j - \bar{P}) \qquad (5.11)$$

This shows that the depletion rate decreases by increasing the water level and increases by increasing the total phosphorus. Figure 59 shows the plot of the depletion rate against water level for different levels of TP. From the figure it appears that for the Central Basin hypolimnion to have a depletion rate of 0.10 mg/L/day, the water level should be approximately 569.5 ft if TP is very low (approximately zero) and 571 ft if TP = .01. Also shown in Figure 59 are the historical values of the depletion rates for the years prior to 1967, as calculated by Charlton (1979), and for 1967 to 1979, as given in the previous chapter. This shows that there was a progressive increase in TP values during the period 1948 to 1970, which is followed by a decrease in TP for 1972 to 1978. The 1979 value suggests an increase in TP. The reality of this inference is clearly seen by noting that the depletion rates for 1962, 1968, 1969, 1970 and 1979 fall between the depletion rates with TP = .02 and TP = .03, while 1948, 1949, 1950, 1951, 1952, 1953 and 1963 depletion rates are below those determined by TP = .01. The inference described here is based on the average conditions (i.e., disregarding the statistical fluctuations of the estimated depletion rates) which are considered next.

Confidence intervals for the estimation of the depletion rates and for testing different hypotheses about their values can be obtained as follows. Let C be the variance-covariance matrix of the parameters of model 5.10, then the submatrix C_1 of C, which contains the elements of C corresponding to the coefficients of $(x_{ij} - \bar{x})$, $(w_j - \bar{w})$ $(x_{ij} - \bar{x})$ and $(x_{ij} - \bar{x})$ $(P_j - \bar{P})$, is the variance-covariance matrix of the coefficients. Hence, for any water level w' and total phosphorus P', the variance of D is:

$$\mathrm{var}(D) = \underline{a}'C_1\underline{a} \qquad (5.12)$$

where $\underline{a}$ is the column vector (1 w'-w P'-P) and $\underline{a}'$ is the transpose of $\underline{a}$. The 1-2α confidence interval for D is

$$D \pm t_{d,\alpha} \sqrt{\underline{a}' C_1 \underline{a}} \qquad (5.13)$$

where $t_{d,\alpha}$ is the tabulated value of the t-distribution with d degrees of freedom and 2α significance level.

This confidence interval for D can be converted into a confidence interval for any of the two independent variables, given the value of the depletion rate and the other independent variable. For example, it is possible to obtain a confidence interval for TP given the depletion rate, the water level and the hypolimnetic temperature. Let D_0 be the value of the depletion rate when the water level is w_0. The estimate P_0 of TP can be obtained for a given value w_0 of the water level by equating D_j to D_0 and solving for P_j in Equation 5.11. The resulting P_j value is the required estimate of TP. The 1-2α confidence interval for P_0 consists of all the values of P_j which satisfy the inequality

$$\frac{(D_0 - D_j)^2}{\underline{a}' C_1 \underline{a}} \leq t^2_{d,\alpha} \qquad (5.14)$$

where in the expression for D_j and $\underline{a}$ the value w_j is set at w_0.

Assuming the rates reported by Charlton and by setting the water level as its observed value, the 95% confidence interval for TP is estimated for the years considered by Charlton (1979). These are shown in Figure 60. In some years, Charlton reported two depletion rates. For these years two confidence intervals are presented and represented by broken curves, and the years with only one depletion rate are represented by a continuous curve. From the figure it can be seen that the confidence intervals shift generally to the right, which indicates an increase in the total phosphorus. The highest increase probably occurred after 1952.

PREDICTION OF ANOXIA IN THE HYPOLIMNION

Model 5.10 can be used to explain the interplay between hypolimnion temperature, water level and total phosphorus in producing anoxic conditions in the lake. Investigating model 5.10 more closely reveals that (i) the depletion rate decreases with increasing water level and increases with increasing TP and (ii) the initial hypolimnetic oxygen concentration is controlled by temperature, TP and by the interaction between temperature and water level. Temperature and TP are inversely associated with the initial oxygen concentration, while the water level effect on the initial O_2 concentration is controlled by temperature also. For a fixed temperature an increase in the water level will result in reducing the initial oxygen concentration, but the degree of the reduction is a function of temperature. To illustrate these points Figure 61 shows the plots of the predicted O_2 concentration from model 5.10 against Julian days for different temperature, water level and TP values. The graph shows that the effect of the water level on O_2 concentration is large for low hypolimnion temperature and is small when the temperature is close to 10°C. Also, for TP = .01 and a stratification period of 110 days the graph shows that anoxia (O_2 concentration = 0) will not occur if the water level is 572 ft or more. On the other hand, anoxia is more likely to occur if the water level is less than 571 ft.

Finally, it is useful to estimate the probability of the occurrence of anoxia for different stratification periods and different phosphorus levels. This is done by noting that the time required for reaching an anoxic condition is estimated by setting $y_{ij} = 0$ and solving of x_{ij} in model 5.10. Hence it is reasonable to call the probability that y_{ij} is less than or equal to zero as the probability of anoxia, PA, which is given by

$$\begin{aligned} PA &= P\{\hat{y}_{ij} \leqslant 0\} \\ &= P\{(\hat{y}_{ij} - \mu_{ij})/\sqrt{\underline{b}'C\underline{b}} \leqslant - \mu_{ij}/\sqrt{\underline{b}'C\underline{b}}\} \\ &= \int_{-\infty}^{-\frac{\mu_{ij}}{\sqrt{\underline{b}'C\underline{b}}}} \frac{1}{\sqrt{2\pi}} e^{-Z^2/2}\, dZ \qquad (5.15) \end{aligned}$$

where C is the variance-covariance matrix of the coefficients of model 5.10, μ_{ij} is the mean value of y_{ij}, $\underline{b}$ is the column vector

$$\{1 \ (x_{ij} - \bar{x}) \ (x_{ij} - \bar{x})(w_j - \bar{w}) \ (T_j - \bar{T}) \ (T_j - \bar{T})(w_j - \bar{w})$$
$$(P_j - \bar{P}) \ (x_{ij} - \bar{x})(P_j - \bar{P})\}$$

and $\underline{b}'$ is the transpose of $\underline{b}$. Note that μ_{ij} is a function of temperature, water level and TP and hence PA is a function of these variables. Since μ_{ij} is not known, its value is estimated by $\bar{y}_{ij}$.

Figure 62 presents the estimated values of PA for a stratification period of 110 days and two levels of phosphorus (TP = .01 and TP = .02). The water levels considered are 570 ft, 571 ft and 572 ft and hypolimnetic temperature is taken as 8°C, 9°C, 10°C, 11°C and

12°C. The figure shows that for TP = .01 and a temperature less than or equal to 10°C, the probability of anoxia is almost 1 for water level equal to 570 ft while for temperature values above 11°C the value of PA is slightly below 1. For water levels of 571 ft and 572 ft, the values of PA increase with temperature. Also, changing TP from .01 to .02 increases the probability of anoxia substantially, especially for water levels above 571 ft.

CONCLUSION

A statistical model is developed for dissolved oxygen concentrations in the hypolimnion of the Central Basin of Lake Erie, using data collected by CCIW during the period 1967 to 1979. Water level hypolimnion temperature and total phosphorus are used as explanatory variables in the model. It was found that the depletion rate is completely independent of temperature and depends only on water level and TP. On the other hand, the initial O_2 concentration in the hypolimnion was found to be a function of temperature, TP and water level. The effect of water level is influenced by the hypolimnetic temperature, i.e., the interaction between temperature and water level. The model is used to show the historical trend in the depletion rate after the removal of the effect of temperature and water level. It is concluded that the increase in depletion is related to the increase in the level of TP. Furthermore, the model is used to estimate the probability of anoxia in the Central Basin as a function of the three explanatory variables. It is concluded that there is always a high chance for the occurrence of anoxia and this chance increases with the increase in the level of total phosphorus. Hence, it is possible to improve the anoxic conditions in the lake by controlling TP loading.

Table 49. Estimate of the Parameters of Model 5.2

Parameter	Estimate	Parameter	Estimate
β_0	1083.0931	γ_5	-0.089
β_1	-1063.618	γ_6	-0.084
γ_1	-0.118	γ_7	-0.066
γ_2	-0.119	γ_8	-0.106
γ_3	-0.106	γ_9	-0.087
γ_4	-0.113	γ_{10}	-0.095

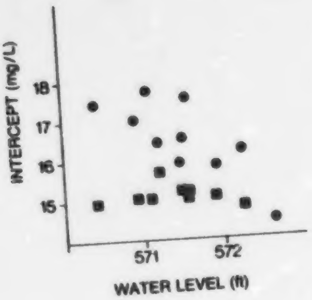


Figure 55. The estimated intercept of the dissolved oxygen regression line against Lake Erie water level. Squares represent Eastern Basin; circles represent Central Basin.

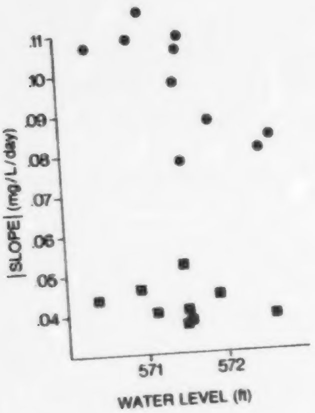


Figure 56. The estimated depletion rate against Lake Erie water level. Squares represent Eastern Basin; circles represent Central Basin.

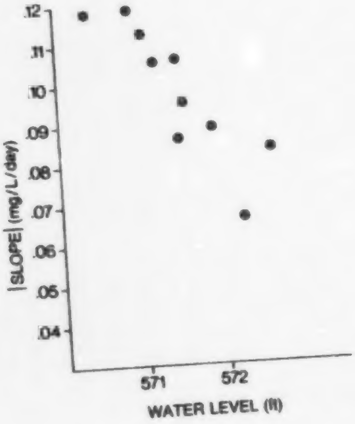


Figure 57. The estimated depletion rate from model 5.2 against Lake Erie water level. Squares represent Eastern Basin; circles represent Central Basin.

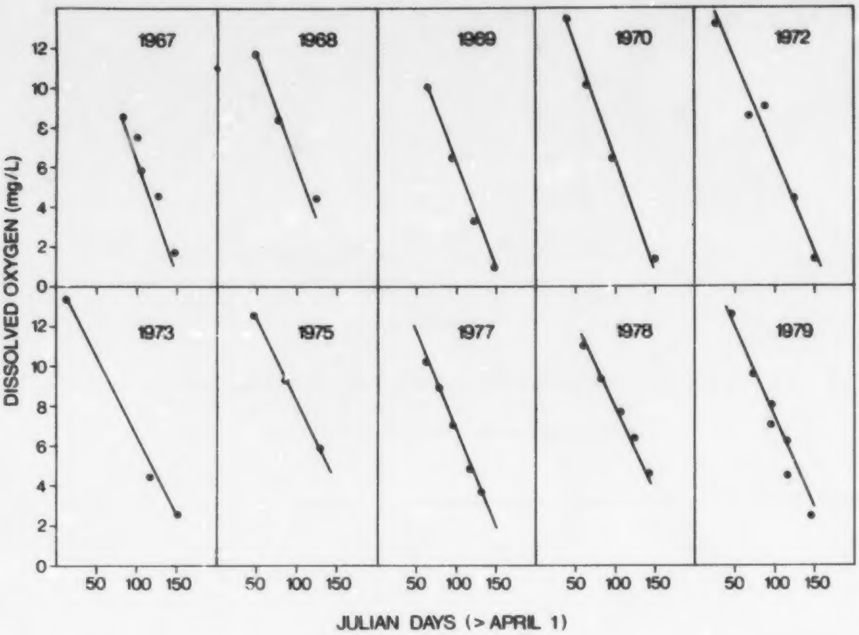


Figure 58. The plot of observed oxygen concentrations and their estimated values from model 5.10 against time in Julian days.

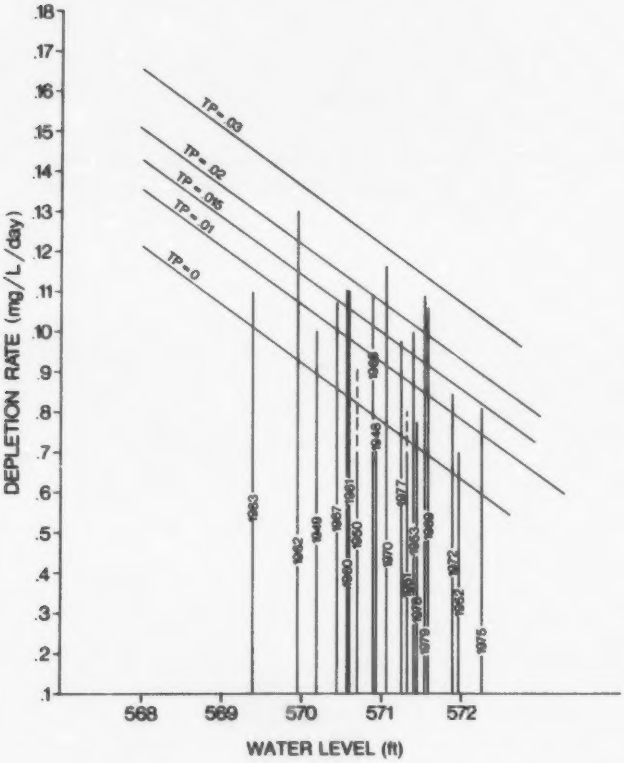


Figure 59. The plot of depletion rate against water level for different levels of TP.

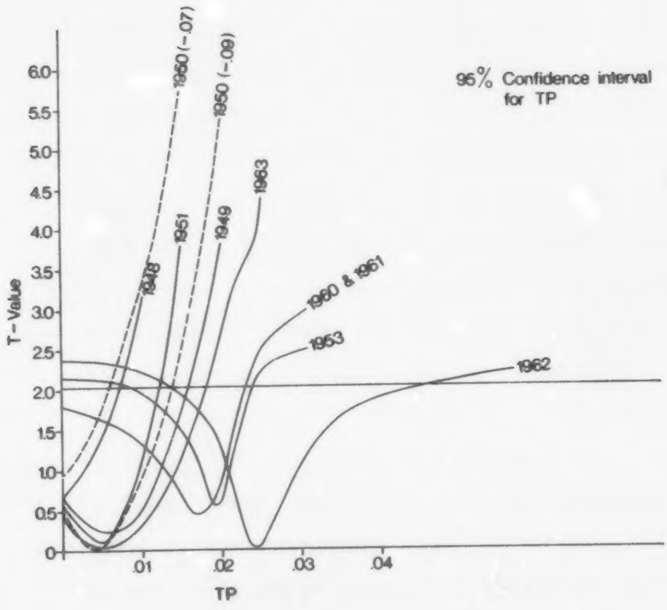


Figure 60. The 95% confidence interval for TP for the years 1948–1962.

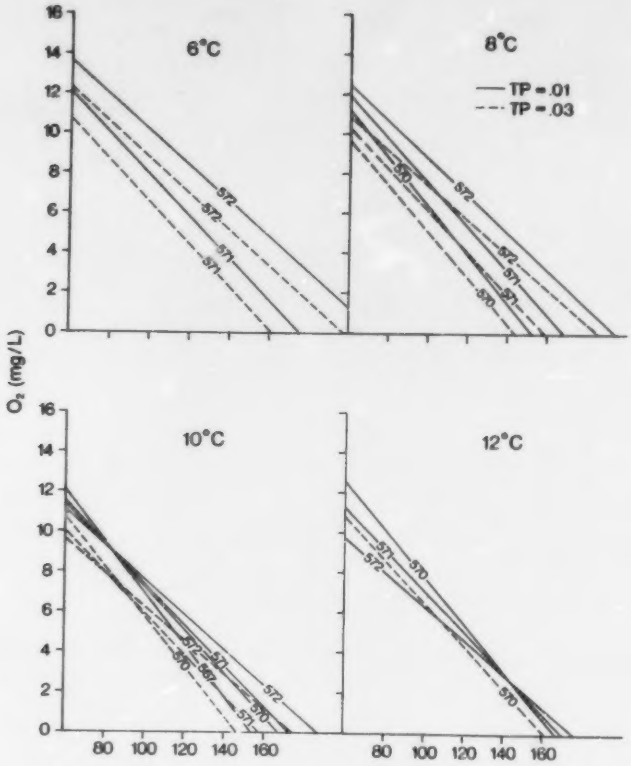


Figure 61. The predicted O_2 concentration from model 5.10 against Julian days for different temperatures, water levels and TP levels.

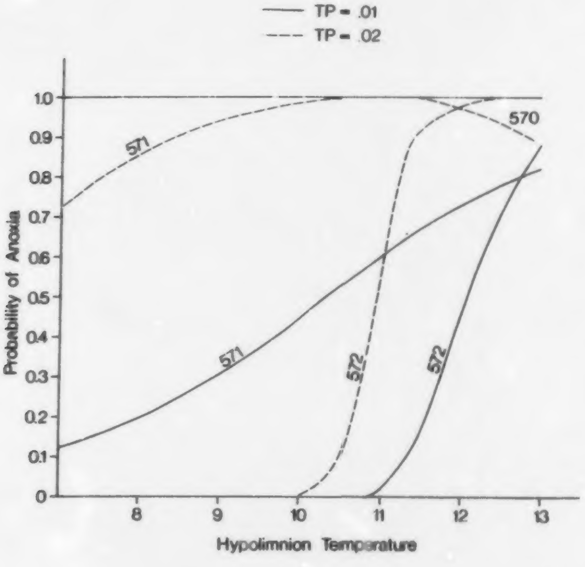


Figure 62. The estimated probability of anoxia for a stratification period of 110 days and different levels for TP and the water level.

CHAPTER 6

Spatial and Temporal Distribution of Total and Fecal Coliform Concentrations, 1966-1970

by S.R. Esterby and A.H. El-Shaarawi

INTRODUCTION

The objective of this chapter is to characterize the spatial and temporal distribution of total coliform and fecal coliform concentrations in Lake Erie as determined using the data from the surveillance program conducted by the Department of the Environment (CCIW, 1966-1976). The total coliform group and fecal coliforms are commonly used as bacteriological indicators of water quality, where it is assumed that their presence indicates that pathogens may be present. For example, the statement of the maximum acceptable level for treated drinking water recommended by Health and Welfare Canada (1979) contains a limit for both total and fecal coliform concentrations, as do the guidelines for treatment of raw water supplies. Recommended limits for bathing waters are based on fecal coliform concentrations (Environment Canada, 1979). Although concentrations obtained from the surveillance program, where a ship is used to collect samples, clearly are not directly relevant to most user-specific guidelines or regulations, the intended inference must be to the bacteriological water quality of the offshore areas as indicated by coliform organisms.

Here a statistical analysis of the data is used to explain the variation in the data in terms of systematic and random components, where the random component takes the form of a probability distribution. The starting point for this analysis arises from the theoretical basis that counts in time or space follow a Poisson distribution (Feller, 1957). It has been shown that, under carefully controlled conditions, bacterial counts on replicate samples of a homogeneous material follow a Poisson distribution (Fisher *et al.*, 1922). However, bacterial concentrations in water samples collected over time and space are generally not well fitted by this simple one-parameter distribution (Pipes *et al.*, 1977; El-Shaarawi *et al.*, 1981). The analysis has thus been separated into several stages.

For the total coliform data, cruise by cruise, the frequency distribution for all the data has been examined and a zonation determined to explain the spatial distribution. Within the limitations of the data as seen from Table 50 and inconsistencies to be discussed later, the data are compared year by year. The relationship between fecal and total coliform concentrations is considered, first using the frequency distribution of fecal coliforms conditional on the total coliform concentrations, and secondly, in terms of the ratio of fecal to total coliform concentration. Since there are years in which fecal coliform concentrations were measured but not total coliform, fecal coliform concentrations for all years are examined separately for consistency over the years. Finally, the relationship between turbidity and total coliform concentrations is briefly considered.

Chapter 6 discusses different statistical methods, presents an analysis of total coliform and fecal coliform data, and then provides a summary and discussion.

STATISTICAL METHODS

Most of the statistical methods used in successive sections are given here. Since the nature of data analysis requires the application of suitable techniques for each particular situation, some further analysis will appear, as appropriate, in the other sections. The total coliform data for 1968 have been taken as the example and are examined in detail in this section.

Empirical Frequency Distributions

The assumption that a random variable R follows a probability distribution denoted by P(R) permits the calculation of the probability with which R takes certain values. Given a set of observations, such as the total coliform concentrations for a set of water samples, the empirical frequency distribution is the analogue of the theoretical probability distribution. By counting the number of times the coliform concentration is equal to each distinct value, the empirical frequency distribution for the set of coliform data is obtained. This can be looked at graphically by means of the frequency histogram, where the area under each bar is equal to the frequency. In this chapter, the empirical frequency distribution is used to select the appropriate probability distribution and to compare the distribution of different sets of data.

Let $r_1, r_2 \ldots r_n$ denote the number of organisms in a standard volume for n water samples; the empirical frequency distribution is then given by

$$f(r) = \frac{\text{Number of } r_i = r}{n} \tag{6.1}$$

Probability Distributions for Bacterial Counts

Summaries of the use of statistics in the analysis of bacterial counts are given by Pipes _et al._ (1977) and El-Shaarawi _et al._ (1981). Three distributions that are commonly used are the

Poisson, negative binomial and lognormal distributions. Let R denote the number of organisms per 100 mL of water. Then, the probability that the number of organisms equals r is given by

$$P(R=r) = e^{-\lambda} \lambda^{r}/r! \tag{6.2}$$

assuming R follows a Poisson distribution or by

$$P(R=r) = \frac{(k+r-1)!}{r!\ (k-1)!} \frac{p^{r}}{(1+p)^{k+r}} \tag{6.3}$$

if R follows a negative binomial distribution. If R follows a lognormal distribution, then $S = \ln R$ has a normal distribution and the tables for the standard normal distribution can be used.

The negative binomial and lognormal distibutions are appropriate when the variation is greater than would be observed if a Poisson distribution fitted the data and the frequency distribution was skewed to the right. For the Poisson distribution, the mean and the variance are equal: $\mu = \sigma^2 = \lambda$. In the case of bacterial concentrations in samples collected over time and space, greater variation can be expected. It has been shown that the negative binomial distribution arises if it is assumed that the mean $\mu = \lambda$ itself varies according to a gamma distribution (Fisher, 1941). The mean and variance of the resulting negative binomial distribution are $\mu = pk$ and $\sigma^2 = p(1+p)k$. Equating the means of the Poisson and negative binomial distributions, $\mu = \lambda = pk$, leads to an expression for the variance of the negative binomial distribution plus a term which represents the excess variance, namely $\sigma^2 = \lambda + \lambda^2/k$.

If a transformation of the data is sought which will stabilize the variance, it can be shown that the logarithmic transformation is appropriate if the variance is proportional to the square of the mean (Brownlee, 1965). Transformation of the data to approximate normality is the reason for using the lognormal distribution for bacterial counts. This latter distribution is often favoured because of the well-known properties of the normal distribution and the ease of the calculations. Complications arise when zero counts are obtained, and the lognormal distribution is completely inappropriate if the number of zeros is too large. Thus situations will occur where one of these two distributions fits the data better than the other.

Fitting Probability Distributions and Estimating Parameters

These two topics, fitting probability distributions and estimating parameters, are not separable but their order of application to a set of data is logically determined. Once a probability distribution has been found which adequately fits the data, statements about the values of the parameters may be made. The test of fit consists of first forming an hypothesis that a particular distribution fits the data and then conducting a test that measures the strength of the evidence against the hypothesis. The choice of the probability distribution may be based upon the characteristics of the empirical frequency distribution or prior knowledge about which probability distributions may be appropriate.

The goodness of fit statistic, X^2 (see for example Snedecor and Cochran, 1967), can be used to test the fit of a probability distribution to a set of data, by grouping the n observations into k classes with the restriction that the classes are formed such that the expected frequency of each class is not too small. Thus

$$X^2 = \sum_{j=1}^{k} (f_j - e_j)^2 / e_j$$

where f_j is the observed frequency in class j, e_j is the expected frequency in class j and X^2 has an approximate chi-square distribution with (k-p-1) degrees of freedom for p = number of parameters in the distribution. The working rule, given by Snedecor and Cochran, that the expectation can be as low as 1 in two classes provided it is not less than 5 for most of the others, was used for the present data.

Since the hypothesis being tested is composite, to calculate the expectation, the parameters of the distribution must be estimated. The maximum likelihood estimates for the parameters p and k of the negative binomial distribution have been used, since for the coliform data analysed here, the method of moments estimates differ considerably from the maximum likelihood estimates. This is discussed later in this section. In testing the fit of the lognormal distribution, the expectation can be calculated directly using the tables of the standard normal distribution.

Denote the mean and variance of $S = \ell nR$, by μ and σ^2, respectively, then $(S-\mu)/\sigma$ has a $N(0,1)$ distribution. By replacing μ and σ^2 by their maximum likelihood estimates,

$$\bar{S} = (\sum_{i=1}^{n} \ell n\ R_i)/n \qquad \text{and} \qquad \hat{\sigma}_S^2 = \sum_{i=1}^{n} (\ell n\ R_i - \bar{S})^2 / (n-1),$$

$P(R \leqslant r) = P(\ln R \leqslant \ln r) = P(S \leqslant s)$ can be obtained.

To test the fit of a Poisson distribution to a set of data it is not necessary to group the data. Fisher's index of dispersion (Fisher _et_ _al_., 1922) D^2 can be used where

$$D^2 = (n-1)\, s^2/\bar{r}$$

and $\bar{r}$ and s are the mean and standard deviation of the n counts $r_1, r_2 \ldots r_n$. The distribution of D^2 is well approximated by a chi-square distribution with $(n-1)$ degrees of freedom.

The sample mean, $\bar{r}$, is the method of moments estimate and the maximum likelihood estimate of λ, the parameter of the Poisson distribution, and thus of the mean and variance of the distribution. Either the relative likelihood function or the confidence limits for λ can be used to provide a measure of the uncertainty involved in estimating λ. The relative likelihood function for λ is given by

$$R(\lambda; r_1, r_2 \ldots r_n) = L(\lambda; r_1 \ldots r_n) \,/\, L(\hat{\lambda}; r_1 \ldots r_n)$$

where

$$L(\lambda; r_1, r_2 \ldots r_n) = \lambda^{\sum_{i=1}^{n} r_i} e^{-n\lambda}$$

and $\hat{\lambda}$ is the maximum likelihood estimate of λ, that is, the value of λ which maximizes L (for a discussion of the use of the likelihood function see Sprott and Kalbfleisch, 1965). Confidence limits for λ can be obtained using the relationship between the Poisson distribution and the χ^2 distribution (Brownlee, 1965). Thus the 1-α confidence limits for λ are given by

$$\tfrac{1}{2}\chi^2_{1-\alpha/2}(2(r.+1)) \text{ and } \tfrac{1}{2}\chi^2_{\alpha/2}(2r.)$$

where $r. = \sum_{i=1}^{n} r_i$ and $\chi^2_p(\nu)$

is the value of the χ^2 variate with ν degrees of freedom for which the cumulative distribution function equals p, and $\chi^2_p(\nu)$ can be obtained directly from tables. If r. is large, the normal approximation to the χ^2 distribution can be used to obtain $\chi^2_p(\nu)$.

The expressions for the first two moments of the negative binomial distribution are $\mu = pk$ and $\sigma^2 = p(1+p)k$. By equating the sample moments to the theoretical moments, the method of moments estimates of p and k are obtained as

$$\tilde{p} = \frac{s^2-\bar{r}}{\bar{r}} \quad \text{and} \quad \tilde{\kappa} = \frac{\bar{r}^2}{s^2-\bar{r}}$$

where $\bar{r}$ and s^2 are the sample mean and variance obtained from $r_1, r_2 \ldots r_n$. Fisher (1941) has shown the conditions for which the method of moments estimates are efficient. High efficiency is obtained if one of the following conditions holds: (1) $\bar{r}$ is small, e.g. $\bar{r} = 0.1$, and $k > 1$; (2) $\bar{r} = 1$ and $k > 3$; (3) $\bar{r} = 10$ and $k > 9$; (4) $p < 1/9$; or (5) $k > 18$. The negative binomial distributions that were fitted to the total coliform data had high values of p and low values of k. Typical of these data is cruise 102 of 1968, for which the efficiency of the method of moments estimates was only about 0.63. In cases where the efficiency of the method of moments is low, the maximum likelihood estimates should be used.

The joint likelihood function for parameters p and k of the negative binomial distribution is

$$L(p,k;r_1,r_2 \ldots r_n) = \left[\prod_{i=1}^{n} \prod_{j=1}^{r_i} (k+r_i-j) \right] p^{n\bar{r}} / (1+p)^{n(k+\bar{r})}$$

The maximum likelihood estimates of p and k are given by $\hat{p} = \bar{r}/\hat{k}$ where $\hat{k}$ is the solution to the equation

$$\sum_{i=1}^{n} \sum_{j=1}^{r_i} \frac{1}{(k+r_i-j)} - n \ln(1+\bar{r}/k) = 0$$

which must be obtained numerically, but can be done readily on a computer. The joint relative likelihood function

$$R(p,k;r_1,r_2 \ldots r_n) = L(p,k;r_1,r_2 \ldots r_n)/L(\hat{p},\hat{k};r_1 \ldots r_n)$$

can be examined to determine plausible values for the pair of parameters. To look at the parameters separately, the maximum likelihood functions for p and k can be used, where

$$L_m(p) = n\bar{r} \ln p - n(\hat{k}+\bar{r}) \ln (1+p)$$

and

$$L_m(k) = \sum_{i=1}^{n} \sum_{j=1}^{r_i} \ln(k+r_i-j) - n(k+r) \ln(1+\hat{p})$$

To interpret the latter two functions, note that the maximum likelihood estimate of the other parameter appears in the expression. Thus the maximum likelihood function of k, for example, gives the plausible values of k, assuming p equals its most plausible value, that is, $p = \hat{p}$.

If R is lognormally distributed with parameters μ and σ^2, $S = \ell nR$ is normally distributed with mean μ and variance σ^2. The maximum likelihood estimates μ and σ^2 are

$$\hat{\mu} = \frac{1}{n} \sum_{i=1} \ell n\, r_i \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (\ell n\, r_i - \hat{\mu})^2$$

The mean and the variance of R are given by $\mu_r = e^{\mu + \frac{1}{2}\sigma^2}$ and $\sigma_r^2 = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$. The method of moments estimates for μ and σ^2 are obtained by equating the first two sample moments of R,

$$\bar{r} = 1/n \sum_{i=1}^{n} r_i \qquad \text{and} \qquad s_r^2 = 1/(n-1) \sum_{i=1}^{n} (r_i - \bar{r})^2$$

to their corresponding theoretical moments and solving for μ and σ^2. This leads to

$$\tilde{\mu} = 2 \log \bar{r} - \tfrac{1}{2} \log (s_r{}^2 + \bar{r}^2)$$

$$\tilde{\sigma}^2 = \log (s_r{}^2 + \bar{r}^2) - 2 \log \bar{r}$$

However, Aitchison and Brown (1957) have shown that the efficiency of the method of moments estimators decreases as σ^2 increases and even for σ^2 as small as 0.5, the efficiency of $\tilde{\mu}$ is 79% and of $\tilde{\sigma}^2$, 31%. As can be seen from the total coliform data discussed later in the chapter, σ^2 is likely much larger than 0.5, since the smallest value observed for the efficient maximum likelihood estimate was $\hat{\sigma}^2 = 0.98$.

If only a test of fit is required estimates of μ and σ^2 are adequate. If estimates of the mean and variance of R are needed, then further calculations are necessary. The method of moments estimates are

$$\tilde{\mu}_r = \bar{r} \quad \text{and} \quad \tilde{\sigma}_r^2 = s_r^2$$

but again these estimates have poor efficiency. Large sample approximations to the maximum likelihood estimates, given by Aitchison and Brown (1957), are as follows

$$\hat{\mu}_r = e^{\bar{s}} \Psi_n (\tfrac{1}{2} s_s^2)$$

$$\hat{\sigma}_r^2 = e^{2\bar{s}} \{\Psi_n (2s_s^2) - \Psi_n (\tfrac{n-2}{n-1} \; s_s^2)\}$$

and $$\Psi_n(t) \doteq e^t \;\{1 - \frac{t(t+1)}{n} + t^2 \;(\frac{3t^2+22t+21}{6n^2})\}$$

where $s_s^2 = \frac{n}{n-1} \; \hat{\sigma}^2$. In some instances, the geometric mean is specified in the statement of a water quality guideline. For example, it is recommended that fecal coliform bacteria in bathing water should not exceed 200 per 100 mL based on a minimum of at least five samples over a 30-day period (Environment Canada, 1979). There is an implicit assumption that fecal coliform count R follows a lognormal distribution. The median of a lognormal distribution with parameters μ and σ^2 is given by e^{μ}. Based on a sample $r_1, r_2 \ldots r_n$, the maximum likelihood estimate of the median is thus $e^{\hat{\mu}}$, but $e^{\hat{\mu}} = e^{\bar{s}}$ = geometric mean, since

$$\bar{s} = 1/n \sum_{i=1}^{n} \ln r_i$$

The lognormal distribution is defined for $r > 0$, and thus zero counts introduce a problem. Two possible solutions are given by Aitchison and Brown (1957). The first is to treat the probability that $R=r$ as the $P(r-\delta<R<r+\delta)$. If a variable $y=\ln(R+1)$ is defined with a probability distribution given by

$$P(Y < 0) = 0$$
$$P(Y = 0) = P(0 < Y < 1) = \int_0^1 f(y)dy$$
$$P(Y \leqslant y) = \int_0^y f(y)dy \text{ for } (y > 0)$$

where $f(y)$ is the probability density function of a normal distribution with mean μ and variance σ^2. The problem of estimation thus reduces to that of a normal distribution censored at the origin. The second solution is to assume that only the non-zero values are lognormally distributed with a proportion δ of zero values and $n-\delta$ of non-zero values. This leads to a Δ-distribution with three parameters, δ = proportion of zero values and μ and σ^2, the parameters of the lognormal distribution. Aitchison and Brown give the maximum likelihood estimators for the parameters and also the mean and variance. A working solution is to add 1 to all the counts and proceed as if R were lognormal. Provided the number of zeros is small and the values of the non-zero observations large, the distortion of the estimates will be small.

Spatial Zonation Procedures

The previous discussion assumes that one probability distribution can adequately describe the bacterial counts for an entire lake. A more useful summary of the bacterial concentrations in the lake can be obtained by dividing the lake into homogeneous zones, that is, zones within which the bacterial concentrations can be described by the same probability distribution but with different distributions for each zone. Since bacterial counts obtained from a homogeneous sample follow a Poisson distribution, it is reasonable to develop a procedure which divides the set of data into subsets, where each subset has a Poisson distribution, but the means are different. One such procedure, based upon the index of dispersion, D^2, is given by El-Shaarawi et al. (1981). An improvement on this is outlined in Esterby and El-Shaarawi (1981), and in El-Shaarawi (1982), a generalization of the procedure to other distributions in the subsets and an alternative stopping criterion are given.

The clustering procedure outlined by Esterby and El-Shaarawi (1981) has been used in this chapter. Assume that there are k groups and let $\lambda_1, \lambda_2 \ldots \lambda_k$ be the means of the Poisson distributions in the k groups. Then an observation r is assigned to group g if $L(r) = P(r;\lambda_g)/P(r;\lambda_j) > 1$ for $j \neq g$, and $P(r;\lambda)$ is the Poisson probability function. That is, the procedure assigns an observation to the subset or group for which the probability that a Poisson variate takes this observed value is the greatest. Thus it is a probability clustering procedure. There are two aspects to this procedure: (1) the optimal allocation for a fixed number of groups and (2) the choice of the number of groups. The procedure contains a number of steps. The number of groups, k, is fixed at its lowest value. Initial values for the parameters $\lambda_1, \lambda_2 \ldots \lambda_k$ are assigned where λ_i is the mean of the Poisson distribution in the ith group and $\lambda_i < \lambda_{i+1}$. The initial values $s_1, s_2 \ldots s_k$ are obtained by dividing the range of the observed values into k equal intervals and setting s_i to the midpoint of the ith interval. Then $c_1, c_2 \ldots c_{k-1}$ are determined where c_i is the value of r which satisfies the equation $L(r) = P(r;\lambda_{i+1})/P(r;\lambda_i) = 1$ and $P(r;\lambda) = \lambda^r e^{-\lambda}/r!$. All of the observations $r_1, r_2 \ldots r_n$ are assigned to the appropriate group by comparison with $c_1, c_2 \ldots c_{k-1}$. If for observation r_j, $r_j > c_i$, r_j is assigned to the i+1 group. The solution to this equation reduces to the simple form

$$c_i = (\lambda_{i+1} - \lambda_i)/(\log \lambda_{i+1} - \log \lambda_i)$$

Once all the observations have been assigned to groups, the parameters $\lambda_1, \lambda_2 \ldots \lambda_k$ are estimated by the mean of the samples in the group. If this is the first iteration, recalculate $c_1, c_2 \ldots c_{k-1}$ and reassign the samples to groups. The assignment to groups and recalculation of the λ's continues until there is no change in group membership on two successive iterations.

The minimum number of groups would normally be 2. For each value of k the indices of dispersion $D_1, D_2 \ldots D_k$ and the corresponding significance levels $\alpha_1, \alpha_2 \ldots \alpha_k$ are calculated. A significant value of D^2, indicating inhomogeneity, in one or more groups indicates that k should be incremented and this procedure of allocation of the samples to groups should be repeated. Thus for any k, the criterion for stopping is met if $\min(\alpha_1, \alpha_2 \ldots \alpha_k)$ is less than the prespecified significance level.

The generalization of the procedure described by El-Shaarawi (1982) allows $P(r_i;\lambda)$ to take forms other than a Poisson distribution. Also the stopping criterion is different and is based on the likelihood ratio

$$\Lambda = \{ \prod_{i=1}^{k} \prod_{j=1}^{n_i} P(r_j ; \lambda_i^{(k)}) \} / \{ \prod_{i=1}^{k+1} \prod_{j=1}^{n_i} P(r_j ; \lambda_i^{(k+1)}) \}$$

where $\lambda_i^{(k)}$ is the value of λ in the ith group when the allocation was assigned assuming k groups, and $\lambda_i^{(k+1)}$ is the value of λ in the ith group when k+1 groups were assumed. Thus it is the ratio of the probability of the observations under the two assumptions, namely, k and k+1 groups. To test the need for the extra group, that is, k+1 rather than k groups, the quantity $-2 \log \Lambda$ can be used, which, for large n, has a χ^2 distribution with one degree of freedom.

As discussed, the grouping procedure does not include the spatial location of the samples. To form spatial zones, the group to which a station belongs is indicated on a lake map. In the initial considerations of zoning methods, it seemed more reasonable to include the geographic location of the station in the grouping procedure itself. A constraint based upon the distance between stations led to too many groups. By comparison with the results of the procedure given by El-Shaarawi et al. (1981), it became apparent that this was due to the discontinuous nature of the zones which occur along the shore. Due to this discontinuous nature, the use of trend surface methods and contouring, suitable for continuous variates (El-Shaarawi and Esterby, 1981), was also ruled out. Thus the interpretation of the zones is simply that regions of the lake, although not necessarily continuous, which have been grouped into one zone exhibit bacterial concentrations that can be considered to come from the same probability distribution and thus, for example, the same mean value can be quoted for all stations in this zone.

An Example: Analysis of 1968 Total Coliform Data

Total coliform data are available from four cruises of Lake Erie in 1968 (Table 50). Sixty-three stations were included in cruise 102, but bacterial analyses were done consistently only at depths of 1 m. The depths and number of stations at which total coliform concentrations were recorded as follows:

Depth	1	7	8.5	9	10	11	12	13	14	15	16
No.	62	32	1	2	5	6	5	4	1	3	3

Depth	17	18	19	20	21	22	27	28
No.	1	3	3	2	2	2	1	1

Since a similar situation appears to hold for other cruises, it was decided that the relationship between concentration and depth would not be investigated and only concentrations at 1 m would be used in the analyses to prevent confounding vertical and horizontal spatial variation.

For each of the four cruises it was found that a negative binomial distribution provided a good fit to the data but a Poisson distribution did not. Because of the high proportion of zeros, 29/62, 13/62, 10/28 and 26/52 for cruises 102, 108, 109 and 112, respectively, the lognormal distribution was not fitted. The observed and fitted frequencies are shown for the range of concentrations from 0 to 18 in Figure 63. Both the Poisson and the negative binomial distributions are discrete and thus their frequencies should appear as bars; curves, however, were drawn so that they could be easily distinguished from the observed frequencies. The observed frequency distributions appear with those of all the other years in Figure 67. The distributions are highly skewed with the modal value r=0.

The tests of fit of the negative binomial distribution to the data of the four 1968 cruises were performed using both the method of moments estimates and the maximum likelihood estimates of p and k in the calculations of the goodness of fit statistic. The observed and expected frequencies, grouped to meet the criterion that two expected frequencies can be as low as 1 provided that the remainder are 5 or more, are given in the top part of Table 52. The value of the goodness of fit statistic, the significance level and the estimates of p and k are given in the bottom part of Table 52. For all four cruises, the method of moments estimate of p is greater than the maximum likelihood estimate and the method of moments estimate of k is smaller than the maximum likelihood estimate. Also, a better fit is obtained for each cruise when the maximum likelihood estimates are used, the most extreme case being cruise 108, where the significance level of 0.08 would indicate some evidence of lack of fit based on the method of moments estimates, compared with a significance level of 0.74 when the maximum likelihood estimates are used. A consistent discrepancy between observed frequencies and expected frequencies based on the method of moments estimates is the overestimation of the number of zeros.

The joint relative likelihood functions for p and k are given in Figure 64 in the form of contours of constant value of the relative likelihood function where $R(p,k) = 0.01, 0.05, 0.50, 0.90$ and 0.99. The relative maximum likelihood functions for p, $R_m(p)$, and for k, $R_m(k)$, are given in Figures 65a and 65b. The joint likelihood contours are triangular, with the range of plausible values of p decreasing as k increases. Because of this, statements about the plausibility of pairs of values of p and k should be based on values of $R(p,k)$ and not taken from $R_m(p)$ and $R_m(k)$. For example, in the case of cruise 102, values of k where $k > 0.32$ or $k < 0.12$ are highly implausible when $p=\hat{p}=32.58$, since $R(\hat{p},k) < 0.01$ for $k > 0.32$ or $k < 0.12$, whereas a larger range of values of k are implausible when $p=97.31$, i.e. $R(p,k) < 0.10$ for $k > 0.19$ and $k < 0.11$. Similar results can be quoted for p. The method of moments estimates $\tilde{p}$ and $\tilde{k}$ have been indicated in Figure 64 for cruise 102 and it can be seen that $R(\tilde{p},\tilde{k})$ is near the 0.50 contour.

Consider negative binomial distributions with parameters equal to the maximum likelihood estimates obtained for these four cruises. In the previous section, the variance of the negative binomial distribution was given as $\sigma^2 = \lambda + \lambda^2/k$, where $\lambda = pk$ is the mean, and the term λ^2/k was called the excess variance relative to the Poisson distribution, for which the variance equals the mean. Thus substituting $\hat{p}$ and $\hat{k}$ in the expression for the variance, the term λ^2/k expressed as a percentage of the total variance equals 97, 98, 97 and 98 for cruises 102, 108, 109 and 112, respectively. To account for this excess variation the zonation procedures have been used.

The grouping procedures, described in the previous section, provide zones for a single cruise, and thus it is necessary to determine how to compare zones from one cruise to another. One possibility is to fix the zones for all cruises at those determined by a chosen reference cruise. The difficulty with this is that the number and location of stations varies from cruise to cruise. Another possibility is to fix the number of zones, and the consequences of this are examined below using the 1968 cruises.

In Table 53, the groups determined for cruise 102 using the probability clustering procedure and both the index of dispersion and the $-2 \log \Lambda$ stopping criterion are compared with the groups determined using the index of dispersion for grouping. For the number of groups k = 2, 3, 4, 5, 6 and 7, the group mean, the number of stations in the group and the value of the index of dispersion and significance level, as obtained from the probability clustering procedure, are given in the top part of the table. The value of -2 log Λ is also given for k = 2,3,4,5, and 6. In the last row of the table, the final groups, obtained using the index of dispersion, are given. When the index of dispersion is used for grouping, four groups are enough to ensure that none of the D^2 values exceeds the 1% critical χ^2 value. If the probability clustering procedure is used, however, seven groups are required if the D^2 stopping criterion is used and six if -2 log Λ is used, where again comparison is with the critical value for the 1% significance level. Comparison of the groups for the two grouping procedures when k=4 shows that the difference lies in the allocation of stations to the first two groups, yielding means of 0.44 and 7.14 with 43 and 14 stations in the first two groups for the probability clustering procedure and means of 0.57 and 7.83 with 45 and 12 stations when the index of dispersion is used for grouping. Note also that as the number of groups is increased from 3 to 6 using the probability clustering procedure, no reallocation is made to groups 1 and 2. Only at k=7 are groups 1 and 2 split.

To assess the consequences of these differences, the probability functions assuming Poisson distributions equal to the group means have been plotted in Figure 66. A characteristic that a grouping procedure should possess and that has not explicitly been incorporated in the present grouping procedures is the separation of the groups. In the present situation, this can be taken to mean that the probability distributions of the groups do not have overlapping ranges of values for which the probability is non-negligible. The overlap can be examined in this figure. In Figure 66a, the probability functions based on the four groups obtained by the probability clustering procedure are shown with the observed frequency distribution given below it. Again for ease in plotting and for visual separation of overlapping probability functions, the dots representing non-zero probability, i.e. points $(r,P(r;\lambda))$ for $r=0,1,2\ldots$, have been joined by lines, but it should be remembered that $P(x;\lambda)=0$ for $r < x < r+1$, where r takes integral values. In Figure 66b, the probability functions based on the first two groups obtained using the index of dispersion are given. Note that groups 3 and 4 are the same for both methods. Since the degree of overlap is essentially the same in Figures 66a and 66b, it appears that the same conclusions would be drawn for the set of data. From Figure 66c, however, it can be seen that splitting of groups 1 and 2 with n=43 and 14 leads to three distributions with non-negligible overlap. Thus, the $-2 \log \Lambda$ stopping criterion is preferred over the D^2 criterion for this set of data. The last question to address is whether the splitting of groups 3 and 4, based on k=4, into four groups with n=1,1,1, and 2 when k=7, is worth the additional computational effort. In view of the sparsity of stations at these high concentrations it was decided that it was not worthwhile. The other three 1968 cruises have been examined in the same way as cruise 102. Again, splitting into a higher number of groups tended to reallocate the stations that occurred in groups 3 and 4 based on k=4. For each of these three cruises, however, there was more overlap of probability functions when D^2 was used for grouping than when the clustering procedure was used. Thus the final strategy chosen for zonation of all the total coliform data was to fix the number of zones at k=4 and to use the probability grouping procedure.

A qualitative comparison of the zones for the 1968 cruises is given in El-Shaarawi et al. (1981). Although the index of dispersion was used for grouping and the number of zones found were 4, 5, 3 and 4 for cruises 102, 108, 109 and 112, respectively, the observations are much the same as those drawn from the zonation summarized in Table 58 and Figure 69. This latter zonation is based on the number of zones fixed at 4 and the probability clustering procedure. The zones of higher concentration form discontinuous regions along the shore with regions of higher concentration spreading across the lake only during cruise 108, which was conducted between July 29 and August 3.

Inference about the Ratio of Bacterial Counts

Since the total coliform count includes bacteria of both fecal and non-fecal origin, and the total coliform count is used to infer the presence of pathogens, it is of interest to know the relationship between fecal coliform and total coliform concentrations. The choice of an indicator, such as total coliform, has always required that the indicator be easier to detect than the pathogen, since it is more numerous. Risk assessment could be greatly improved if a stable ratio of pathogen to indicator could be determined, and thus inference about the ratio of two bacterial counts is important (El-Shaarawi and Pipes, 1982). El-Shaarawi and Pipes consider inference about the ratio of the means of two independent Poisson distributions and of the medians of two lognormal distributions.

Inference about the ratio of fecal coliform to total coliform may be required when data are acquired under several different situations: (1) the fecal and total coliform concentrations are measured on the same water samples as independent measurements; (2) as in (1) but as dependent measurements, i.e. fecal coliform identifications made from the total coliform plate; and (3) separate sets of fecal and total coliform data, from the same water, are available for estimation of the ratio. The latter situation is clearly less favourable than (1) or (2).

Application of the results given by El-Shaarawi and Pipes (1982), assuming total and fecal coliform measurements are dependent, leads to the following. Assume F=fecal coliform concentration and O=non-fecal coliform concentration are independently Poisson distributed with means λ_f and λ_o, respectively. Inference about the ratio $\rho=\lambda_f/\lambda_o$ can then be based on the conditional distribution of F given T=F+O, where T is the total coliform concentration. On the basis of $f_1,f_2...f_n$ and $t_1,t_2...t_n$, where $o_i=t_i-f_i$, the maximum likelihood estimate for ρ, using the conditional likelihood function, is just

$$\hat{p} = \sum_{i=1}^{n} f_i \;/\; \sum_{i=1}^{n} o_i$$

Clearly, the ratio $\rho=\lambda_f/\lambda_o$ cannot be estimated by direct application of the results of El-Shaarawi and Pipes if T and F are independent, because O less than zero is a possibility. For T and F independent, the ratio $p_1 = \lambda_f/\lambda_t$, assuming T is Poisson-distributed with mean λ_t, can be estimated by this method, yielding

$$\hat{p}_1 = \sum_{i=1}^{n} f_i \;/\; \sum_{i=1}^{n} t_i$$

When the total and fecal coliform concentrations are determined on the same sample, however, this corresponds to the situation of matched pairs. If T and F are i[illegible]ndent and Poisson-distributed, a more appropriate model is to [illegible] that the mean of T_i is a function of a parameter specific to the pair and one associated only with the total coliform concentration, and similarly for F. If the functional form of the mean is multiplicative, then

$$\mu_{t_i} = \beta_i \lambda_t \quad \text{and} \quad \mu_{f_i} = \beta_i \lambda_f$$

where μ_{t_i} is the mean of T_i, μ_{f_i} is the mean of F_i and β_i is the parameter specific to the pair. For $\rho = \lambda_f/\lambda_t$ this yields

$$P(t_i) = \theta_i^{t_i} \frac{e^{-\theta_i}}{t_i!} \qquad\qquad P(f_i) = (\rho\theta_i)^{f_i} \frac{e^{-\rho\theta_i}}{f_i!}$$

where $\theta_i = \beta_i \lambda_t$ and thus $\mu_{f_i} = \beta_i \lambda_f = \beta_i \lambda_t \rho = \theta i \rho$. But this is the special case of the power series distribution considered by Kalbfleisch and Sprott (1973) and leads to the conditional likelihood

$$\prod_{i=1}^{n} \binom{t_i + f_i}{f_i} \rho^{\sum_{i=1}^{n} f_i} / (1+\rho)^{\sum_{i=1}^{n} (t_i + f_i)}$$

and $\hat{\rho} = \sum_{i=1}^{n} f_i / \sum_{i=1}^{n} t_i$. This is the same result as obtained above using the results of El-Shaarawi and Pipes (1982) and assuming T and F are independent. If $\mu_{t_i} = \beta_i + \lambda_t$ and $\mu_{f_i} = \beta_i + \lambda_f$, then estimation of $\rho = \lambda_f/\lambda_t$ is more difficult.

ANALYSIS OF TOTAL COLIFORM DATA

Total coliform data are available for 16 cruises of Lake Erie conducted between August 1966 and August 1970; the distribution between years is poor, however, and the number and locations of the stations are highly variable from cruise to cruise. The analysis of the data described in this section follows the order in which the methods were described previously.

The observed frequency distribution is given for each of the 16 cruises in Figure 67, where the chronological order is obtained by reading down the left side of the page, then the right side and lastly, the second page. An arrow with a number beside it indicates an observed value which is beyond the range of values shown on the histogram. The frequency distribution is given on two scales in the adjacent histograms. For the 1967, 1968 and 1969 cruises, the left histogram covers the range of observations and thus most of these histograms consist of one spike on the left side of the histogram with much smaller bars scattered along the range. These plots have been included to emphasize the extreme skewness of the data and to permit comparison of the less skewed 1968 data with those of 1967 and 1969. The histograms on the right of each pair, which cover the range of concentration from 0 to 14 or 0 to 16, show that the modal value is 0 and that even over this range, the distribution is very skewed. The two histograms on the right side of the pair are based on the frequency for each interval of 5, given by the cross-hatched bars. For the 1966 and 1970 cruises, the left histogram is the same as described above, but the right histogram is the frequency distribution of the natural logarithm of the concentration. As stated previously, the reason for using the logarithmic transformation is to meet the assumption of normality. The resulting frequency distribution of the 1970 cruise is much more symmetric than that of the untransformed data. Nevertheless, the frequency distribution is still skewed in the case of cruise 111 in 1966 and appears to be bimodal. Histograms are given, based on a grouping interval of 0.5 and 1.0 $\log_e R$. The large spike on the interval [0,0.5] corresponds to the zero counts. For the

1966 data, 1 has been added to all the counts before taking the logarithm.

From the observed frequency distributions, it can be seen that the data separate into two inherently different types of data sets: one group of data from 1967, 1968 and 1969 and the other from 1966 and 1970. There is a common shape of the frequency distribution for each cruise in 1967 and 1968. Most of the concentrations are in the range from 0 to 100 coliforms per 100 mL with many 0 values recorded and marked skewness even in the range from 0 to 16, with very few concentrations exceeding 100. This feature occurs even though the sampling pattern is variable, as can be seen from the station locations shown in Figure 68. Note, however, that the frequency distributions have been shown for cruises 107 in 1967 and 102, 108, 109 and 112 in 1968 using a range of only 0 to 100 in the histogram on the left of each pair, whereas a range of 0 to 1000 was used for all other cruises in these years.

The 1969 data have been recorded in a different manner from the data of all other years. Many 0 counts occur in 1966, 1967 and 1968, but none occurs in the 1969 data. Instead, <1, <2 and <3 have been recorded. To put 1969 on the same basis as the other years, these indeterminate values have been changed to 0. It is clear that this is reasonable for values given as <1. The <2 and <3 values have been set to 0 because it appears that these arise from the multiplication of 1, in the <1, by the factor used to convert the concentration to the units of number per 100 mL, followed by rounding to the nearest integer if necessary. On the basis of the data with zeros, it can be seen from Figure 67 that the 1969 cruises are similar to those in 1967 and 1968, although cruise 110 in 1969 has fewer zeros relative to other low counts than do the other cruises in this group.

The cruises in 1966 and 1970 showed more than occasional values exceeding 100 coliforms per 100 mL. In 1966, there was the spike of zero values as for 1967 and 1968, but a peculiar distribution can be seen in Table 51 where it is compared with that of cruise 102 of 1968 and cruise 107 of 1970. Thus the bimodal character of the transformed data is due to the range of values from 0 to about 100, with a distribution similar to those of the 1967 and 1968 cruises and the rest, a distribution similar to that of the 1970 cruise.

The conclusion from this is that the data for 1967, 1968 and 1969 can probably be examined for differences between years, but that comparisons over the years 1966 to 1970 can only be qualitative. The differences in the data of 1966 and 1970 from 1967, 1968 and 1969 cannot be ascribed to different sampling patterns, but may be due to changes in analytical procedures i.e. medium and filters (B.J. Dutka, personal communication).

It was shown previously that the total coliform counts from the 1968 cruises were fitted by negative binomial distributions. Because the 1967 and 1969 cruises also have a large number of zero counts, the Poisson and lognormal distributions were ruled out, and the negative binomial fit to the data was tested. It was found that the negative binomial distribution does not fit these data. By examining the three cruises of 1967, given as examples in Table 54a, it can be seen that in each case, the number of values in the right tail of the distribution far exceeds that expected for a negative binomial distribution. The range of values for the 1968 cruises is 0 to 260, while it is 0 to 150 000 for the 1967 and 1969 cruises. Furthermore, the value of 260 formed a separate group when the clustering procedure was applied to cruise 108 of 1968. The extremely high values clearly do not belong to the same distribution as the rest of the concentrations. To exclude moderately high values, the 1967 and 1969 cruises were compared with those of 1968. Values were excluded if they exceeded 200 and if their exclusion provided groups comparable with those of 1968 when the clustering procedure was applied. This ad hoc procedure was considered adequate, since it was not clear whether total coliform counts in large lakes would look like the 1967-1968-1969 group or the 1966, 1970 group. However, if the 1967-1968-1969 type of data occurs in other cases, a more formal procedure would be justified.

The results from the tests of fit of the negative binomial distribution to the cruises, with high values removed, are given in Table 55. The test was performed using both the method of moments and the maximum likelihood estimates for p and k, the parameters of the negative binomial distribution. To see which values have been excluded and the estimates of p and k, refer to Table 56. The details of the test are given for cruises 101, 103 and 105 of 1967 in Table 54b. The maximum likelihood estimates of p and k have been used for the tests given in Table 54b.

The fit of the negative binomial distribution to cruises 101, 103 and 105 of 1967 is much better once the high values have been removed (compare Tables 54a and 54b). Lack of fit is still indicated for cruise 101, but this is due to a higher than expected number of counts equal to 1, not to lack of fit in the upper tail as was observed when all the data were used. For cruises 103 and 105, the individual terms of the χ^2 goodness of fit statistic are not as large as this for any particular value and overall the fit is adequate for these two cruises. Despite the large numbers of observations, n=161 and 165, for the latter two cruises, the observed frequency distributions are somewhat irregular.

From the rest of the 1967 and 1969 cruises in Table 56, it can be seen that when the maximum likelihood estimates for p and k are used, the negative binomial distribution fits the data adequately for all cruises except cruise 109 in 1967. As with cruise 101, this is mostly due to more concentrations equal to 1 than expected. For cruises 101 to 115 of 1967, all four 1968 cruises, and cruise 103 of 1969, the method of moments estimates of p and k are markedly different from the maximum likelihood estimates, and for each of these, except cruise 109 of 1967, where both sets of estimates produce highly significant results, the conclusion drawn about the fit would differ depending upon which pair of estimates of p and k was used. Except in cruise 101 of 1967, fit is poorer in all these cases when the method of moments estimates are used than when the maximum likelihood estimates are used. Since it is known that the method of moments estimates are not efficient for such low values of k, the conclusion is reinforced, based on the 1968 data, that the maximum likelihood estimators should be used.

The observed frequency distributions showed that the distributions for 1967, 1968 and 1969 are extremely skewed. In addition, by fitting a theoretical distribution to the data, it was shown that with two exceptions, a negative binomial distribution fitted the data for a single cruise. This is quite remarkable in view of the large variation in the number of stations sampled and in the number of zeros observed (Table 56). When a series of data sets, such as the cruise data used here, can be fitted by the same model with varying parameter values, the sets can be compared by following the values of the parameters. The maximum likelihood estimates of p and k have been plotted for 1967, 1968 and 1969 in Figure 68; the only conclusion that seems to follow, however, is that the parameters vary erratically. Little more can be expected from examination of all of the data for one cruise because of the variable number and location of sample stations.

The data for cruise 107 of 1970 are well fitted by a lognormal distribution, as can be seen from Table 57. As discussed above, the data of cruise 111 of 1966 would appear to be better fitted if the logarithmic transformation was applied as for the 1970 data. However, 23 of the 92 observations are zeros and cannot be transformed directly. One was added before transformation, that is, the transformation was taken as log(R+1), and the fit tested. As can be seen from Table 57, the lognormal distribution does not fit these data due to the large number of zeros and presence of two other peaks in the observed frequency distribution. The model called the Δ-distribution (Aitchison and Brown, 1957), which assumes that the population consists of a proportion δ of 0 values with the non-zero values following a lognormal distribution, would be more appropriate here. The extra computation has not been performed, since it has been adequately established that the 1966 data are not like those of any other years.

The probability clustering procedure, using the number of groups fixed at 4 and with high values excluded, has been applied to all 16 cruises and the results are given in Table 58 and Figure 69. The excluded values have been called zone 5 and are given for each cruise as the eleventh column of Table 58. The seasonal pattern described for the 1968 cruises is at least partially confirmed for the other years.

The earliest cruises are cruise 101, May 30 to June 8, 1967; 102, May 17-22, 1968; and 103, May 30 to June 4, 1969. In all three cases, most of the offshore stations appear in zone 1 and the other zones occur around Cleveland and as discontinuous groups along the south shore. In the 1968 and 1969 cruises, zone 2 appears along the shore at the northeast end of the lake and the zones of higher concentration appear in the Western Basin, and in 1968, in the north western part of the Central Basin as well. Note that the means for the corresponding zones are similar: 0.10, 0.44 and 0.19 in zone 1; 7.50, 7.14 and 10.31 in zone 2; 40.00, 43.33 and 51.40 in zone 3; and 87.50, 79.50 and 130 in zone 4, with the order being cruise 101 in 1967, 102 in 1968 and 103 in 1969.

In the second 1968 cruise, conducted July 29 to August 3, the zones of higher concentration had spread across the lake in the eastern end. This could not be confirmed in the Central Basin due to a lack of samples. Since there are more cruises in 1967, the second cruise was conducted earlier, and it can be seen that by June 20 to 29 (cruise 103), the zones of higher concentration are spreading across the lake. In fact, for cruises 103, 105, 107, 113 and 115 of 1967, there are bands of higher concentration spreading across the lake between the northern and southern shores at variable locations. Cruises 109, August 21 to 30, and 111, September 11 to 17, do not appear to have these bands, but the sampling in the Central Basin was less dense for these cruises. Cruise 103, June 20 to 29, shows the highest concentrations in 1967 as does cruise 108, July 29 to August 3 in 1968. Since no cruise was conducted in late June of 1968, however, it cannot be determined whether the peak concentrations occurred later in 1968 or whether they were missed. Although the concentrations in 1966 and 1970 are so different from those of 1967, 1968 and 1969, qualitatively the spatial patterns are similar to those of 1967, 1968 and 1969 for the same period of the year. Thus the 1966 cruise, 111 conducted between August 8 and 14, and the 1970 cruise, 107, conducted between July 28 and August 2, have the bands of higher concentration spreading across the lake.

The last cruises, conducted in October or November of 1967, 1968 and 1969, appear to be returning to the spring pattern, that is, fewer stations in the zones of higher concentration occur in the centre of the lake. In 1967, however, the October cruise still shows a band across the lake which was not detected in the late August and the September cruises.

Thus, in a qualitative sense the spatial pattern has been characterized for three periods of the year. For any particular cruise the results are quantitative, but a description of the zones which will apply to all the years of data cannot be quantitative due to the differences between the data from year to year, the small number of cruises in all years except 1967 and the inconsistent sampling pattern. The means for zones 1, 2, 3, and 4 combined, for zone 1 and for zone 2 can be compared for 1967, 1968 and 1969 in the top part of Figure 68. The peak concentrations in cruise 101 of 1967 and 108 of 1968 can be seen in the plot of the overall means and in zones 1 and 2 for 1967, but only in zone 1 of 1968. The discontinuous regions of higher concentration along the shores occur in all cruises but not necessarily at the same location, with the exceptions of the vicinity of Cleveland, where high concentrations are always observed at some of the stations, and Erie, where high concentrations are observed most of the time. To show more clearly the difference between the zones in the spring and summer, zone 2 has been taken as an example, and the stations in zone 2 during cruise 101 May 30 to June 8 and cruise 103 June 20 to 29 in 1967, are given in Figure 70. The locations of the stations in zones 3, 4 and 5 for the 1967 cruises are shown in Figure 71. Furthermore, the stations that were excluded from the calculations required to obtain zones 1 to 4, and called here zone 5, occur, with four exceptions, in the vicinity of Cleveland, Toledo, Erie, the mouth of the Detroit River, or the mouth of the Grand River in Ontario (Table 59). The four exceptions are also given in the table and can be located on the appropriate maps in Figure 69. Note that between 1967 and 1969, values exceeding 1000 coliforms/100 mL occur only near either Toledo, Cleveland or the mouth of the Grand River. The explanation of the bands of higher concentration (usually zones 2 and 3) which spread across the lake likely rests with the patterns of water circulation, since transport from the shore must be involved. Either of the following two features of the Central Basin circulation (Hamblin, 1971; Simons, 1976) could be involved: (1) circulation cells from the flow along the north shore which extend throughout the basin or (2) summer surface currents which predominantly move away from the north shore.

ANALYSIS OF FECAL COLIFORM DATA

Fecal coliform data are available for 15 cruises between 1967 and 1975, but only for 11 of these were total coliforms measured (Table 50). The relationship between total and fecal coliforms is of interest for several reasons, and thus the emphasis of the section is on this relationship rather than on the fecal coliform concentrations alone. In the guidelines for the treatment of raw water used as the source of drinking water, Health and Welfare Canada (1979) recommends that three criteria be met. Each of these contains a statement about the maximum concentration allowed for both fecal coliform and total coliform concentrations. The guidelines for treated drinking water also require that none of the coliform organisms detected be fecal coliforms. In this context, the empirical frequency distribution of fecal coliforms given the total coliform concentration is examined. A second consideration is whether total coliforms provide an adequate indicator of pathogens, since the total coliform count includes organisms from both fecal and non-fecal origin. In this context, the ratio, which has been discussed previously, is calculated.

For all 15 cruises, the frequency distribution of fecal coliforms consists of a high proportion of zeros (0.52 to 0.87), with the remaining observations spread over a large range, and in most cases, the frequency of these observations is one (Table 60). Higher fecal coliform concentrations occurred during the 1967 cruises than during the other years. The locations of stations with fecal coliform concentrations greater than 200 have been given in a footnote to Table 60. With a few exceptions, again as for total coliforms, the high concentrations occur in the vicinity of Cleveland, Erie, Toledo or the Grand River. For cruise 110 in 1969, the observations quoted as less than a particular value have been converted to zeros, and the same argument given for total coliforms holds.

As a first step in examining the relationship between total and fecal coliforms, fecal coliforms have been plotted against total coliforms using the stations from all the 1967 and 1968 cruises for which the total coliform concentration did not exceed 100 bacteria per 100 mL (Figure 72, top). The entire range of values is 0 to 51 000 for total coliforms and 0 to 12 000 for fecal coliforms. The same plot, limiting total coliforms to less than 1000, is also given in Figure 72, and the large values which have not been plotted are listed below, with total coliforms in the first row and fecal coliforms in the second.

1000	2600	6000	6300	8300	11 000	12 000	22 000	51 000
400	340	880	5000	400	1 600	3 500	250	12 000

The only conclusion that can be drawn from these plots is that in the low range of total coliform concentrations, 0 to 30, there is no relationship between the two concentrations, and although higher fecal coliform concentrations are observed when higher total coliform concentrations occur, the variance also becomes very large.

The statement of the criterion for drinking water given above, i.e., that none of the total coliform detected be fecal coliform, when put into the form of a probability statement requires a conditional probability distribution. Since total and fecal coliform concentrations from the same samples are available in the present set of data, the observed frequency distribution of fecal coliforms conditional on the total coliform concentration can be examined. This is done by obtaining the frequency distribution of the fecal coliform concentration for which the total coliform concentration is the same. Let R_t be the total coliform concentration and R_f be the fecal coliform concentration. Then the conditional frequency function of R_f given R_t, $f(r_f/R_t=r_t)$, is given by the number of stations with both fecal coliform concentration equal to r_f and total coliform concentration equal to r_t divided by the number with total coliform concentration equal to r_t. These frequency distributions are given in Figure 73 for cruises 103, 105 and 107 of 1967 and 102, 108 and 112 of 1968 for the range of the total coliform concentration restricted to 0 to 20, and for all the 1967 and 1968 data in Table 61. It becomes clear that a large amount of data is required to determine the form of the conditional frequency function. On the basis of Table 61, it is evident that low fecal coliform concentrations are not observed when high total coliform concentrations occur. For the present data, however, large numbers of observations are available only for low total coliform concentrations, and thus above $R_t>40$, little can be said about $f(R_f/R_t)$. For low R_t,

$f(R_f/R_t)$ is skewed with the mode at 0, and $f(R_f/R_t)$ appears to become less skewed as R_t becomes larger. The latter observation is characteristic of a number of skewed distributions. For example, as the mean of the Poisson distribution increases the distribution function becomes more symmetric. Although the form of the conditional distribution has not been determined here, the present discussion is of importance because it gives the initial step for consideration of guidelines such as those given above. The importance of knowledge concerning the form of the frequency distribution in the formulation of regulations and guidelines has been discussed by Esterby (1982).

Note also that for low concentrations, the fecal coliform concentration exceeds the total coliform concentration in a few cases, a result that is possible if independent analyses are performed on the same sample but not if the fecal coliform concentration results from identifications made on the total coliform plate. This was discussed previously and is relevant to the assumption made in the estimation of the ratio.

The zones determined in the previous section, using all of the total coliform concentrations, have been used to study the spatial distribution of fecal coliforms. Since fecal coliforms were not measured at all of the stations at which total coliforms were measured, the subset of common stations has been identified for each zone. The summary of the fecal coliform concentrations for the common stations is given in Table 62 and the summary of the total coliform concentrations in Table 63. The examination of the relationship between total and fecal coliform concentrations given earlier in this section was based on pooled sets of data. The zonation allows the data to be examined once the spatial component has been eliminated.

The fecal coliform concentration is plotted against the total coliform concentration for each zone of the 1967 and 1968 cruises in Figure 74 and the mean fecal coliform concentration versus the mean total coliform concentration in Figure 75. In Figure 74, the vertical scale is the fecal coliform concentration and the horizontal scale is the total coliform concentration, and the number of occurrences of a particular pair of values is plotted instead of the symbol (•) if the number is more than one. Again, the lack of relationship at low concentrations is apparent except for zone 2 of cruise 112 in 1968.

As mentioned previously, the ratio, ρ, of the fecal coliform concentration to the total coliform concentration can be estimated by the ratio of the corresponding sample means. This has been done using only the stations in common for zones 1, 2 and 3 of the cruises in 1967, 1968 and 1969 and also using the total coliform means based on the entire data set. The results are given in Table 64 and the estimated ratios, $\hat{\rho}$, based on the common stations are plotted in Figure 76.

The ratio for zone 1 varies erratically for the 1967 cruises, largely due to the occurrence of a fecal coliform concentration greater than the total coliform concentration at one or more stations in the zone. It can be seen from Figure 74, that this occurs for cruises 105, 107, 111, 113 and 115, cruises for which the ratio in zone 1 is larger than that of cruise 103 and 109. The 45 degree line has been drawn on the plots in Figure 74 for which the fecal coliform concentration exceeds the corresponding total coliform concentration at one or more stations. Note also that the station in zone 1 of cruise 111 with a fecal coliform concentration of 40 has been excluded from the calculations. In zone 2 there are three occurrences of a fecal coliform concentration exceeding the total coliform concentration (cruises 105, 113 and 115), and again, as was the case for zone 1, these are the cruises with a higher ratio. The question arises whether these stations should be excluded. From Table 60, it can be seen that the frequency distributions for fecal coliforms are highly skewed. The frequency distributions of fecal coliform concentrations within zone 2 for the 1967 cruises are also highly skewed, as are the conditional frequency distributions (Table 61). Thus, the chance of obtaining a relatively high fecal coliform concentration in a zone is not negligible and therefore these stations were not removed. Thus the ratios for cruises 105, 113 and 115 ($\hat{\rho} > 0.20$) exceed those of cruises 103, 107, 109 and 111 ($\hat{\rho} < 0.10$) because for the former cruises occasional high fecal coliform concentrations were observed.

In 1968, the ratio is lower for cruise 108 than for cruises 102 and 112. Since the ratio could be estimated for only one cruise in 1969, nothing can be said about this year. There is one consistent feature of the variation of the ratio for both the 1967 and 1968 data. The ratio is low during the cruise of peak total coliform concentration, a time when the zones of higher concentration spread across the lake. In 1968, this pattern occurs only in cruise 108 and thus there is a difference in spatial distribution which agrees with the difference in the ratios. However, in 1967 there are no zone 2 bands across the lake in cruise 101 and again in August and September, in cruises 109 and 111, but the pattern of absence of these zones in the middle of the lake in the early and late cruises is not observed as it was in 1968. Furthermore, the differences in ratios between the 1968 cruises are accompanied by a different relationship between fecal and total coliform concentrations, which can be seen from Figure 74. There appears to be no relationship between fecal and total coliform concentrations for cruise 102 of 1968. There is an increase in the variance of fecal coliform concentrations as total coliforms increase for cruise 108 as well as an increase in fecal coliform concentration with the increase in total coliforms for cruise 112. This is in marked contrast to the reason for the difference between the cruises of 1967.

These differences may be due to the differences in sampling pattern for fecal coliforms in 1967 and 1968 (Figure 77). In 1967, samples were taken at stations in the Western Basin, and near Cleveland, Erie and the mouth of the Grand River in Ontario, although there was considerable variation from cruise to cruise. In 1968, samples were taken along most of the shore, although again there was considerable variation from cruise to cruise. Thus the 1967 fecal coliform concentrations are from stations known to be near on-shore sources of bacterial pollution, while the 1968 concentrations are from stations near the shore whether next to on-shore sources or not. Note that only the 1969 cruise permits comparison of fecal and total coliform concentrations for the entire lake. Thus the conclusion concerning the seasonal pattern for the ratios awaits further data.

The range of plausible values for ρ can be obtained from the relative conditional likelihood function of ρ. Previously, the conditional likelihood function $L_c(\rho)$ was given and from this it follows that

$$L_c(\rho) \propto \rho^{f.}/(1+\rho)^{t.+f.}$$

where for n pairs of independent fecal and total coliform concentrations $(f_1,t_1),(f_2,t_2),\ldots,(f_n, t_n)$ and $f. = \sum_{i=1}^{n} f_i$ and $t. = \sum_{i=1}^{n} t_i$. The relative likelihood function is given by

$$R_c(\rho) = \frac{L_c(\rho)}{\max\limits_{\rho} L_c(\rho)}$$

and has been plotted for zone 2 of cruises 103 and 105 of 1967 and 102 and 108 of 1968 in Figure 78. There is overlap of the range of plausible values, i.e. $R_c(\rho) \geqslant 0.50$, for all four cruises shown in Figure 76, which means that the ratios are not really different. Again the most reasonable conclusion to draw from these data is that they are inadequate in relation to $R_c(\rho)$; the appropriateness of the assumption that both fecal and total coliform concentrations are Poisson-distributed is unknown.

Examples of the maps which could be obtained for the fecal coliform concentrations are given in Figure 79. The fecal coliform concentration, for the total coliform zones, is shown as a dot (•) if the concentration is 0 and as a number if the concentration is greater than 0. For cruises 103 and 105 of 1967, it can be seen that the conclusions about the fecal coliform concentrations are restricted to the vicinity of urban areas or river mouths where, it has been shown above, total coliform concentrations are high. Thus any inference about the ratio would not apply to the whole lake.

SUMMARY AND DISCUSSION

The observed frequency distributions of the total coliform concentrations for each cruise show that the 1966 and 1970 data are inherently different from the 1967, 1968 and 1969 data. In the latter three years, there are many zero concentrations and for each cruise the distribution is very skewed. Most of the concentrations are in the range 0 to about 200 organisms per 100 mL, and the concentrations in this range can be fitted by a negative binomial distribution in 2 of the 14 cruises. The lack of fit for the two cruises is due to a broader peak in the observed distribution than in the fitted negative binomial distribution. In 1967 and 1969, there are a few higher concentrations in each cruise which do not belong to the same population as the other concentrations and come from stations characteristically near large urban areas or river mouths. The 1970 data are completely different. The total coliform concentrations range from 20 to 2700 and are well fitted by a lognormal distribution. The 1966 data look like a combination of the other two types but are not fitted by either a lognormal or a negative binomial distribution.

It is evident that considerable variation still exists in the concentrations on the reduced range, 0 to 200, for any one cruise. Since the period of time taken to complete each cruise is small, the source of variation must be spatial differences. This has been borne out by the results of the application of a probability clustering procedure to all the cruises between 1966 and 1970. Despite the differences from year to year, a seasonal pattern of the spatial distribution emerges. Early in the year the zone of lowest concentration covers most of the lake, with the zones of higher concentration forming discontinuous regions along the shore. Later, the concentration peaks and zones 2, and sometimes 3, spread in bands across the lake. The discontinuous regions of higher concentration along the shore are still present and, in fact, remain during all cruises, although the position is not necessarily the same. In the fall, the pattern appears to begin to return to the earliest spatial distribution, but there are differences between the years, largely due to the existence of zone 2 bands across the lake late in 1967. The discontinuous regions along the shore can be attributed to sources on shore and this confirms what one would expect. The bands that spread across the lake, however, must be due to circulation patterns of the lake. The spread of zones 2 and 3 across the lake in the 1968 July-August cruise was reported by El-Shaarawi et al. (1981). Only when all the available data had been analysed did it become apparent that this pattern repeated itself and that the location of the band is variable.

Reitz (1973) reports total coliform concentrations determined at water intake stations along the south shore of Lake Erie. He notes that these concentrations obtained at a considerable distance from the shore are lower than beach concentrations but higher than his observed average of 2 coliform bacteria per 100 mL about 2 miles offshore. He attributes high concentrations to local shore processes and the eastward shoreline current along the south shore. Rao and Burnison (1976) report the 1970 total coliform data analysed here and state that certain areas such as Cleveland, Erie and the northwestern part of the Central Basin have higher concentrations (>500) than the main body of the lake. They also attribute the high concentrations in the northwest area of the Central Basin to flow patterns. Thus the observations of this chapter are in agreement with these general comments. The more specific description of the spatial and seasonal pattern of Lake Erie of this chapter, however, provides considerably more information than previously available. Furthermore, the zones of higher concentration spreading as bands across the lake, which have been attributed to lake currents, are a new observation, since these must be due to the circulation cells and not just the shoreline currents. The consequence of these bands, at the very least, is to produce higher offshore concentrations than expected.

Warmer water temperatures result in a better survival rate of coliform bacteria (Reitz, 1973). To determine whether the temperature is related to the presence of these bands, the total coliform concentrations and temperature were plotted for the region around a band of zone 2 stations for cruises 103 and 107, and, using the same region as plotted for 103, for cruise 109, all of which are 1967 cruises. Note that zone 2 was not present in the middle of the lake for cruise 109. It was apparent that warmer temperatures are not responsible for the presence of bands due to either generally warmer temperatures in the lake or warmer temperatures in the zone 2 band relative to the neighbouring stations. The lake is warmer during cruises 107 and 109 than 103, and bands occur during cruises 103 and 107 but not during cruise 109. The differences in temperatures for the regions plotted during any one of the three cruises consist of a gradient from north to south shore and not differences between stations in zones 1 and zones 2 and 3.

The data are inadequate for year to year comparisons due to the inherent differences in the concentrations, the irregular sampling pattern and an inadequate number of cruises for most years. These comments apply also to the fecal coliform concentrations. It should be noted again that in 1969, total and fecal coliform concentrations recorded as less than some number have been converted to zeros to conform to the method of recording the data used in other years.

The fecal coliform concentrations for a cruise consist mostly of zeros with a few non-zero values scattered over a large range. Again high concentrations, up to 12 000 organisms per 100 mL, were observed in the vicinity of large urban areas or river mouths, and the variable sampling pattern, which sometimes included these vicinities, accounts for the large disparity in the maximum fecal coliform concentration observed on different cruises.

The relationship between fecal and total coliform concentrations has been considered in several ways. Plots of all the data where fecal and total coliform concentrations were obtained at the same station and within the four zones, determined using all the total coliform data, indicate that at total coliform concentrations below about 30 there is no relationship between the fecal and total coliform concentrations, with the exception of zone 2 in cruise 112 of 1968. Over a broader range fecal coliform concentrations are high when total coliform concentrations are high, but there is enormous variability. Looking at this relationship in terms of the ratio of fecal to total coliform mean concentration by zone leads only to the conclusion that at peak total coliform concentration the ratio is low, but the data are inadequate for conclusions about seasonal effects. A preliminary conclusion about the conditional frequency distribution of fecal coliforms, given the total coliform concentration, is that like the unconditional distribution, it is skewed to the right. The significance of this with respect to water quality guidelines is that for a given total coliform concentration, the chance of obtaining a non-zero fecal coliform concentration is greater than would be observed for a symmetric distribution.

Since an indicator organism is intended to detect occurrences of pathogens, the performance of total coliforms as an indicator organism of fecal contamination can be examined by comparison of the fecal and total coliform concentrations at common stations. For this purpose the conditional frequency function is more useful than the ratio because the number of times a zero fecal coliform concentration is observed when a non-zero total coliform concentration is found will indicate when the total coliform concentration is detecting organisms of a strictly non-fecal origin. For the 1967 and 1968 cruises, when only stations along the shore or in the vicinity of a large urban area or river mouth were sampled, 129 of the 215 stations, with non-zero total coliform concentration, had zero fecal coliform concentration. From Table 61, it can be seen that 126 of the 129 occurred at total coliform concentrations less than 50. This can be looked at in more detail in Figure 74. For example, in zones 1, 2, 3 and 4 of cruise 103 in 1967, the proportion of stations with zero fecal coliform concentration but non-zero total coliform concentration are 10/11=0.91, 9/13=0.69, 3/5=0.60, 0/1=0, respectively. Thus, it appears that the indicator organism is giving fewer false positives at higher concentrations.

One final consideration with respect to the contribution of non-fecal coliforms to the total coliform count is to compare the turbidity and the total coliform concentration. The Western Basin is more turbid than the other basins due to the high silt input from the heavily silt-laden streams tributary to this basin, and many common bacteria in the soil are included in the total coliform count.

The total coliform concentration has been plotted against turbidity for the 1968 cruises including only the stations in the Western Basin (Figure 80). From these very limited data there is no indication that total coliform counts increase with turbidity.

All of the available 1968 data have also been examined. Total coliform concentration is plotted against turbidity for each cruise in Figure 81. From the figure, Table 65 and Figure 82, where the mean turbidity and mean total coliform concentration are given by zone using again the zones based on all the total coliform concentrations, it can be seen that at low turbidity (<4) there is no relationship between total coliform concentration and turbidity. There may be a relationship over a larger range but there are not enough higher turbidity readings in 1968 to draw conclusions. This would be worthwhile pursuing further, including a factor to separate turbidity due to silt and other sources and examining the basins separately. Such an analysis, if a relationship is found, should show that soil bacteria are contributing to the total coliform concentration when the fecal coliform concentration is zero but the total coliform concentration is not.

All of these results are based on the application of statistical methods which are intended to identify sources of variability and to quantify them. Some well-known methods have been used as well as some relatively new methods. The simple concept of the observed frequency distribution has helped to separate sets of data, which although apparently collected using the same analytical methods, are very different. This is made more quantitative by fitting probability distributions to the data. The comparison of the maximum likelihood estimates of the parameters of the negative binomial distribution with the method of moments estimates has led to the recommendation to use the former estimates for data such as the 1967, 1968 and 1969 total coliform counts. This result and the nature of the joint likelihood function for these parameters are results of considerable importance for situations where the negative binomial distribution is being used.

Even though the data had many shortcomings, the detection of seasonal patterns in the spatial distribution of the total coliform concentrations by means of the zones illustrates the power of the probability clustering technique described here. Note that for more favourable data sets the arbitrary restriction to four zones could be removed and the procedure left to choose the number of zones. However, for the present data, it was considered more important to make years comparable. Finally, estimation of a ratio of bacterial counts and conditional frequency distributions are given. The latter have been shown to be more relevant to several questions in the comparison of fecal and total coliform concentrations.

Table 50. **Summary of the Number of Stations where Total Coliforms, Fecal Coliforms or Both Were Measured for All Available Surface Data***

		Number of stations		
Year-cruise	Dates	Total coliforms	Fecal coliforms	Both
66-111	Aug. 8-14	92	--	--
67-101	May 30 - June 8	150	--	--
67-103	June 20-29	166	47	47
67-105	July 10-19	170	57	57
67-107	July 31 - Aug. 10	99	31	31
67-109	Aug. 21-30	97	30	30
67-111	Sept. 11-17	97	27	27
67-113	Oct. 2-9	95	33	30
67-115	Oct. 24-30	75	21	21
68-102	May 17-22	62	35	34
68-108	July 29-Aug. 3.	62	35	33
68-109	Aug.31-Sept. 2	28	--	--
68-112	Nov. 5-10	52	30	23
69-103	May 30-June 4	59	--	--
69-110	Oct. 15-20	73	72	72
70-107	July 28-Aug. 2	47	--	--
74-104	Aug. 21-25	--	30	--
75-101	Apr. 3-10	---	32	--
75-106	June 24-29	--	12	--
75-107	Aug. 5-10	--	31	--

* Only concentrations from samples at a 1-m depth were used, since at no other depth were coliform densities consistently measured.

Table 51. Observed Frequency Distribution, f(r), of Total Coliform Concentrations, r, for Cruise 111 of 1966, 102 of 1968, and 107 of 1970

Cruise 111 - 1966		Cruise 102 - 1968		Cruise 107 - 1970	
r	f(r)	r	f(r)	r	f(r)
0	23	0	29	20	1
1	4	1	9	62	1
3	4	2	5	80	1
5	3	3	2	82	1
0	1	4	2	90	1
9	1	5	3	110	2
10	14	6	1	130	2
14	2	7	1	140	1
17	1	9	1	180	1
20	3	10	1	200	2
25	1	11	1	210	1
30	4	12	1	220	1
50	5	18	1	230	1
60	1	35	1	260	1
70	1	36	1	300	2
100	1	59	1	310	2
110	2	79	1	340	1
140	1	80	1	360	1
170	1		— 62	380	1
180	1			420	3
190	2			440	1
210	1			460	2
220	1			500	1
230	1			530	1
240	1			550	1
250	1			560	1
290	2			610	1
310	1			680	1
320	2			740	1
340	1			790	1
380	1			880	1
470	1			890	1
590	1			1000	1
760	1			1100	2
5200	1			1200	1
	— 92			1500	1
				2700	2
					— 47

Table 52. Goodness of Fit Tests of the Negative Binomial Distribution to the 1968 Total Coliform Data Using Both the Method of Moments and Maximum Likelihood Estimates for p and k

1968 Cruise 102						1968 Cruise 108					
Method of moments			Maximum likelihood			Method of moments			Maximum likelihood		
r	f	e	r	f	e	r	f	e	r	f	e
0	29	34.07	0	29	30.49	0	13	20.29	0	13	12.28
1	9	5.33	1	9	5.98	1	4	5.27	1,2	7	8.79
2,3	7	5.14	2-4	9	7.89	2-4	6	7.69	3,4	3	5.26
4-76	7	5.04	5-8	5	5.02	5-8	8	5.46	5-7	7	5.56
8-16	5	5.14	9-17	5	5.47	9-14	9	5.02	8-12	8	6.37
17-44	2	5.05	18-43	2	5.09	15-24	9	5.08	13-19	8	5.94
45-61	1	1.02	44-60	1	1.04	25-43	5	5.25	20-29	4	5.47
62	1	1.24	61	2	1.01	44-90	7	5.11	30-45	4	5.09
						91-113	0	1.03	46-87	7	5.18
						114	1	1.80	88-112	0	1.04
									113	1	1.02

1968 Cruise 109						1968 Cruise 112					
Method of moments			Maximum likelihood			Method of moments			Maximum likelihood		
r	f	e	r	f	e	r	f	e	r	f	e
0	10	13.40	0	10	10.66	0	26	30.88	0	26	26.32
1-5	11	6.94	1,2	5	4.53	1-4	12	8.53	1,2	6	7.31
6-33	5	5.40	3-9	6	5.50	5-13	5	5.09	3-6	9	5.60
34-52	1	1.03	10-35	5	5.27	14-47	7	5.08	7-16	5	5.64
≥ 53	1	1.24	36-52	1	1.01	46-69	1	1.05	17-47	4	5.06
			≥ 53	1	1.03	≥ 70	1	1.37	48-68	1	1.04
									≥ 69	1	1.03

Goodness of Fit Test

	68-102		68-108		68-109		68-112	
	MM	ML	MM	ML	MM	ML	MM	ML
Goodness of fit statistic	7.04	4.63	12.77	5.19	3.32	0.15	3.01	2.61
Degrees of freedom	5	5	7	8	2	3	3	4
Significance level	0.22	0.47	0.08	0.74	0.19	0.99	0.40	0.63
Estimate of p	41.10	32.58	68.07	41.59	40.90	34.61	62.67	42.62
Estimate of k	0.16	0.20	0.26	0.43	0.19	0.27	0.13	0.18

r,f,e - Denote bacterial concentration, observed frequency and expected frequency, respectively.
MM - Denotes method of moments estimates.
ML - Denotes maximum likelihood estimates.

Table 53. Comparison of Grouping Procedures Using Total Coliform Data

	Group mean							n							Index of dispersion (D^2)							(a)
k	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	-2 log Λ
Probability clustering procedure																						
2	2.09	57.8						57	5						332 (<<.001)	33.5 (<.001)						982.601
3	0.44	7.14	57.8					43	14	5					46.6 (>.30)	27.7 (.01)	33.5 (.001)					128.014
4	0.44	7.14	35.5	72.7				43	14	2	3				46.6 (>.30)	27.7 (.01)	.01 (>.90)	3.86 (>.20)				85.406
5	0.44	7.14	35.5	59.0	79.5			43	14	2	1	2			46.6 (>.30)	27.7 (.01)	.01 (>.90)	0 -	0.60 (> 70)			8.113
6	0.44	7.14	35.0	36.0	59.0	79.5		43	14	1	1	1	2		46.6 (>.30)	26.7 (.01)	0 -	0 -	0 -	0.01 (>.90)		0.014
7	0.24	3.46	10.83	35.0	36.0	59.0	79.5	38	13	6	1	1	1	2	29.0 (>.70)	7.29 (>.80)	4.33 (.50)	0 -	0 -	0 -	0.01 (>.90)	
Index of dispersion (b)																						
4	0.57	7.83	35.50	72.67				45	12	2	3				59.6 (.06)	21.96 (.04)	0.01 (>.90)	3.86 (>.10)				

(a) El-Shaarawi, 1982, Table 2.
(b) El-Shaarawi et al., 1981, Table 4.

Table 54a. Comparison of the Fit of the Negative Binomial Distribution to Some of the 1967 Total Coliform Data Using All of the Data for the Cruise and a Subset with High Values Removed

Cruise 101 (n=150)				Cruise 103 (n=166)				Cruise 105 (n=170)			
r	f	e	χ^2,(α)	r	f	e	χ^2,(α)	r	f	e	χ^2,(α)
0	119	120.03	0.01	0	65	74.94	1.32	0	77	89.05	1.63
1-3	14	8.01	4.48	1	21	9.40	14.32	1	20	8.55	15.32
4-12	7	5.86	0.22	2	5	5.28	0.02	2,3	26	7.95	40.98
13-16	2	1.30	0.38	3,4	12	6.65	4.30	4-6	17	6.35	17.88
≥17	8	1.02	47.94	5-7	10	6.24	2.27	7-10	6	5.19	0.13
				8-11	11	5.53	5.42	11-16	10	5.10	4.70
			53.02	12-17	8	5.69	0.94	17,18	1	1.32	0.08
			(<<0.001)	≥18	34	2.21	457.88	>>19	13	1.20	116.50
							486.45				197.21
							(<<0.001)				(<<0.001)

Table 54b. Comparison of the Fit of the Negative Binomial Distribution to Some of the 1967 Total Coliform Data Using a Subset with High Values Removed

Cruise 101 (n=146)				Cruise 103 (n=161)				Cruise 105 (n=165)			
r	f	e	χ^2,(α)	r	f	e	χ^2,(α)	r	f	e	χ^2,(α)
0	119	119.62	0.00	0	65	65.94	0.01	0	77	78.39	0.02
1	13	6.09	7.84	1	21	15.21	2.20	1	20	19.93	0.00
2-4	4	6.79	1.15	2	5	9.18	1.90	2	17	11.86	2.23
5-12	4	6.08	0.71	3	6	6.69	0.07	3	9	8.40	0.04
13-41	3	5.38	1.05	4,5	8	9.67	0.29	4	8	6.42	0.39
42-61	1	1.02	0.00	6,7	8	6.98	0.15	5	1	5.13	3.32
≥62	2	1.02	0.94	8,9	8	5.44	1.20	6,7	11	7.75	1.36
				10-12	8	6.36	0.42	8-10	3	7.83	2.98
			11.72	13-16	3	6.35	1.77	11-14	5	6.54	0.36
			(0.02)	17-21	1	5.83	4.00	15-20	8	5.55	1.08
				22-28	9	5.79	1.78	21-34	3	5.07	0.85
				29-37	8	5.01	1.78	35-43	1	1.11	0.01
				38-53	5	5.29	0.02	≥44	2	1.02	0.94
				54-94	4	5.19	0.27				
				95-119	1	1.05	0.00				13.58
				≥120	1	1.00	0.00				(0.19)
							15.89				
							(0.26)				

r,f,e,χ^2 - The number of coliforms per 100 mL, observed frequency, expected frequency, value of $(f-e)^2/e$. The sum of these latter terms and the corresponding significance level, α, are given at the bottom of each χ^2 column.

n - Equals the number of observations.

Note: Maximum likelihood estimates of p and k were used.

Table 55. Test of Fit of the Negative Binomial Distribution to the Total Coliform Data for 1967, 1968 and 1969

			Method of moments			Maximum likelihood		
Year	Cruise	n	χ^2	Degrees of freedom	α	χ^2	Degrees of freedom	α
1967	101	146	3.76	3	0.29	11.72	4	0.02
	103	161	15.56	13	0.28	15.89	13	0.26
	105	165	12.83	9	0.18	13.58	10	0.19
	107	98	28.90	5	< 0.001	11.08	6	0.09
	109	95	100.01	4	< 0.001	25.01	7	< 0.001
	111	94	4.91	4	0.30	1.90	4	0.75
	113	91	36.03	4	< 0.001	8.53	6	0.20
	115	73	7.14	5	0.21	6.78	5	0.24
1968	102	62	7.04	5	0.22	4.63	5	0.47
	108	62	12.77	7	0.08	5.19	8	0.74
	109	28	3.32	2	0.19	0.15	3	0.99
	112	52	3.01	3	0.40	2.61	4	0.63
1969	103	54	9.25	5	0.10	2.50	5	0.77
	110	68	11.76	8	0.17	11.82	8	0.17

Table 56. Estimates of the Parameters of the Negative Binomial Distribution for the Total Coliform Data of 1967, 1968 and 1969

Year	cruise	n	Values excluded	Method of moments $\bar{p}$	Method of moments $\bar{k}$	Maximum likelihood $\hat{p}$	Maximum likelihood $\hat{k}$	Mean	Variance	Median	Number of zeros
1967	101	146	230,520,610,1000	54.32	0.04	45.15	0.05	2.35	129.97	0	119
	103	161	270,330,530,600,8300	41.35	0.25	42.92	0.24	10.13	429.00	1	65
	105	165	300,340,500,800,11 000	17.06	0.23	14.44	0.27	3.93	70.92	1	77
	107	98	6300	46.87	0.10	25.20	0.19	4.86	232.52	0	50
	109	95	750,22 000	93.35	0.07	29.36	0.21	6.19	583.98	1	41
	111	94	900,1000,150 000	10.10	0.18	7.90	0.22	1.78	19.72	0	57
	113	91	500,6000,12 000,51 000	74.97	0.07	31.05	0.18	5.54	420.74	0	47
	115	73	500,2600	29.41	0.21	30.46	0.20	6.07	184.54	1	36
1968	102	62		41.10	0.16	32.58	0.20	6.58	277.03	1	29
	108	62		68.07	0.26	41.59	0.43	17.95	1239.92	8.5	13
	109	28		49.90	0.19	34.61	0.27	9.36	476.31	2	10
	112	52		62.67	0.13	43.62	0.18	7.87	500.82	0.5	26
1969	103*	54	360,700,4200,4900,7000	60.34	0.21	76.37	0.17	12.74	781.48	0.5	27
	110*	68	210,480,520,540,770	41.00	0.36	43.40	0.34	14.66	615.78	3	18

*Indeterminate values have been converted to zeros.

Table 57. Test of Fit of Lognormal Distribution to Total Coliform Data of Cruise 111 of 1966 and Cruise 107 of 1967

Cruise 111, 1966				Cruise 107, 1970			
Class limits $\log_e R$	Observed frequency f	Expected frequency e	$\frac{(f-e)^2}{e}$	Class limits $\log_e R$	Observed frequency f	Expected frequency e	$\frac{(f-e)^2}{e}$
≤ 0.5	23	14.27	5.33	≤ 4.5	5	4.05	0.23
[0.5,1.0)	4	5.62	0.47	[4.5,5.0)	5	5.11	0.00
[1.0,1.5)	4	6.73	1.11	[5.0,5.5)	6	7.8[illegible]	0.42
[1.5,2.0)	4	7.64	1.73	[5.5,6.0)	8	9.32	0.19
[2.0,2.5)	15	8.23	5.57	[6.0,6.5)	11	8.66	0.64
[2.5,3.0)	3	8.41	3.48	[6.5,7.0)	6	6.26	0.01
[3.0,3.5)	8	8.16	0.00	≥ 7.0	6	5.79	0.01
[3.5,4.0)	5	7.51	0.84				
[4.0,4.5)	2	6.55	3.16			χ^2_4 =	1.49
[4.5,5.0)	4	5.43	0.38				
[5.0,5.5)	8	4.27	3.27			α =	0.83
[5.5,6.5)	10	5.43	3.85				
≥ 6.5	2	3.75	0.82				
		χ^2_{10} =	30.00				
		α =	0.001				

[a,b) indicates that the concentration is $\geq$a and <b.

Table 58. Zones 1 to 5 for Total Coliforms by Cruise from 1966 to 1970

		Zonation by likelihood ratio										
		Mean for zone				Number in zone					Zones 1 to 4	
Cruise	Date	1	2	3	4	1	2	3	4	Zone 5	Min.	Max.
66111	Aug. 8-14	5.20	56.33	251.18	606.67	56	15	17	3	5200	0	760
67101	May 30-June 8	0.10	7.50	40.00	87.50	132	10	2	2	230,520,610,1000	0	100
67103	June 20-29	1.16	19.68	61.20	160.00	113	37	10	1	270,330,530,600,8300	0	160
67105	July 10-19	0.21	3.81	14.86	39.50	97	48	14	6	300,340,500,800,11 000	0	60
67107	July 31-Aug. 10	0.36	6.44	34.33	100.00	70	23	3	2	6300	0	100
67109	Aug. 21-30	0.52	6.41	37.00	220.00	71	17	6	1	750,22 000	0	220
67111	Sept. 11-17	0.33	6.29	35.00	-	76	17	1	-	900,1000,150 000	0	35
67113	Oct. 2-9	0.18	4.22	33.33	180.00	57	27	6	1	500,6000,12 000,51 000	0	180
67115	Oct. 24-30	0.33	6.50	40.00	70.00	48	18	6	1	500,2600	0	70
68102	May 17-22	0.44	7.14	43.33	79.50	43	14	3	2		0	80
68108	July 29-Aug. 3	1.52	13.23	43.42	260.00	27	22	12	1		0	260
68109	Aug. 31-Sept. 2	1.33	14.25	33.50	110.00	24	4	2	1		0	110
68112	Nov. 5-10	0.98	14.67	44.00	150.00	40	9	2	1		0	150
69103*	May 30-June 4	0.19	10.31	51.40	130.00	31	16	5	2	360,700,4200,4900,7000	0	140
69110*	Oct. 15-20	2.02	23.50	63.57	130.00	46	14	7	1	210,480,520,540,770	0	130
70107	July 28-Aug. 2	144.35	439.47	1022.22	2700.00	17	19	9	2		20	2700

* Indeterminate values have been converted to zeros.

Table 59. Total Coliform Concentrations and Location of the Stations in Zone 5

Year cruise	Vicinity of concentration				Concentration	Location
	Toledo	Cleveland	Erie	Grand River Ontario		
1966,111					5200	Eastern Basin, middle
1967,101		230,520,610,1000				
103		270,330,530,8300		600		
105	800	11 000	500	300	340	Eastern Basin, S.E. end
107		6300				
109	750	22 000				
111	150 000	900		1000		
113	51 000	12 000	500	6000		
115		2600		500		
1969,103		4200,4900,7000			360 700	Long Point Detroit River
110		210,520			480	Point Pelee
					540,770	Detroit River

Table 60. Observed Frequency Distributions for Fecal Coliform Data, 1967-1975

r	Frequency	f(r)
67-103		
0	34	0.723
1	1	0.021
2	2	0.043
4	2	0.043
6	1	0.021
8	1	0.021
28	1	0.021
34	1	0.021
68	1	0.021
92	1	0.021
120	1	0.021
400	1	0.021
	n = 47	1.000
67-105		
0	42	0.737
2	8	0.140
8	1	0.018
10	2	0.035
16	1	0.018
90	1	0.018
300E	1	0.018
1600C	1	0.018
	57	1.000
67-107		
0	26	0.839
2	2	0.065
14	1	0.032
44	1	0.032
5000C	1	0.032
	31	1.000
67-109		
0	26	0.867
4	1	0.033
32	1	0.033
250C	1	0.033
250W	1	0.033
	30	1.000
67-111		
0	22	0.815
1	1	0.037
2	1	0.037
4	1	0.037
40	1	0.037
400G	1	0.037
	27	1.000
67-113		
0	17	0.515
2	4	0.121
4	1	0.030
6	1	0.030
8	1	0.030
30	1	0.030
34	1	0.030
150	1	0.030
190	1	0.030
880G	1	0.030
3 500C	1	0.030
12 000T	1	0.030
	33	1.000
67-115		
0	14	0.667
1	1	0.048
2	2	0.095
6	1	0.048
17	1	0.048
42	1	0.048
340C	1	0.048
	21	1.000
68-102		
0	22	0.629
1	4	0.114
2	2	0.057
3	2	0.057
6	1	0.029
18	1	0.029
19	1	0.029
53	1	0.029
54	1	0.029
	35	1.000
68-108		
0	27	0.771
2	2	0.057
4	1	0.029
6	1	0.029
7	1	0.029
15	2	0.057
24	1	0.029
	35	1.000
68-112		
0	16	0.533
2	4	0.133
4	2	0.067
5	1	0.033
6	1	0.033
9	1	0.033
13	1	0.033
14	1	0.033
72	1	0.033
90	1	0.033
350C	1	0.033
	30	1.000
69-110		
0	58	0.806
1	6	0.083
2	2	0.028
3	1	0.014
4	1	0.014
6	1	0.014
8	1	0.014
10	1	0.014
48	1	0.014
	72	1.000
74-104		
0	20	0.667
1	1	0.033
2	1	0.033
3	1	0.033
6	1	0.033
9	1	0.033
10	1	0.033
13	1	0.033
21	1	0.033
34	1	0.033
53	1	0.033
	30	1.000
75-101		
0	25	0.781
1	4	0.125
27	1	0.031
39	1	0.031
210L	1	0.031
	32	1.000
75-106		
0	8	0.667
1	1	0.083
3	2	0.167
16	1	0.083
	12	1.000
75-107		
0	16	0.516
1	1	0.032
2	1	0.032
5	1	0.032
7	1	0.032
10	1	0.032
43	1	0.032
76	1	0.032
83	1	0.032
100	1	0.032
110	2	0.065
180	1	0.032
190	1	0.032
310E	1	0.032
430C	1	0.032
	31	1.000

Note: The letter r indicates fecal coliform concentration.
The letters beside the concentrations give the location of the station, where C, E, G, L and T represent the vicinity of Cleveland, Erie, Grand River, Long Point and Toledo respectively.

Table 61. Conditional Frequency Distribution $f(R_f/R_t)$ for Fecal Coliforms, R_f, Given Total Coliforms, R_t, Based on 1967 and 1968 Data

		Number				Number				Number	
R_t	R_f	$R_t \cap R_f$	R_t	R_t	R_f	$R_t \cap R_f$	R_t	R_t	R_f	$R_t \cap R_f$	R_t
0	0	112	118	9	0	3	4	[40,50)	0	6	12
	1	1			4	1			2	1	
	2	4							4	1	
	40	1		10	0	8	8		16	2	
									15	1	
1	0	25	29	11	0	1	2		42	1	
	2	4			1	1					
								[50,100)	[0,10)	4	9
2	0	17	21	12	0	6	8		[10,20)	1	
	1	2			2	1			[20,30)	2	
	2	2			8	1			[30,40)	0	
									[40,50)	0	
3	0	8	11	13	0	1	1		[50,60)	2	
	2	1									
	6	1		14	2	1	2	[100,500)	[0,25)	1	9
	10	1			6	1			[25,50)	4	
									[50,100)	3	
4	0	12	15	15	7	1	1		[100,150)	1	
	2	2									
	3	1		16	0	4	7	[500,1000)	[0,50)	2	6
					2	1			[50,100)	1	
5	0	10	13		13	1			[100,200)	1	
	2	2			14	1			[200,300)	1	
	3	1							[300,400)	1	
				17	9	1	1				
6	0	5	8					[1000,5000)	340	1	2
	1	1		18	2	1	2		400	1	
	4	1			8	1					
	5	1						[5000,10 000)	400	1	3
				[20,30)	0	8	13		880	1	
7	0	1	4		2	3			5 000	1	
	1	1			4	1					
	2	1			52	1		[10 000,25 000)	250	1	3
	8	1							1 600	1	
				[30,40)	0	4	12		3 500	1	
8	0	7	8		2	2					
	4	1			4	2		51 000	12 000	1	1
					6	1					
					10	1					
					14	1					
					18	1					

Columns 3, 7 and 11 give the number of stations with total coliform concentration equal to R_t and fecal coliform concentration equal to R_f.
Columns 4, 8 and 12 give the number of stations with total coliform concentration equal to R_t.
[a,b) indicates that the concentration is $\geq a$ and $<b$.

Table 62. Summary of Fecal Coliform Concentrations at Stations for Which Both Total and Fecal Coliforms Were Measured

		Mean for zone				Number in zone				Zone	Zones 1 to 4	
Cruise	Date	1	2	3	4	1	2	3	4	5	Min.	Max.
67103	June 20-29	0.17	1.23	7.00	34.00	24	13	5	1	68,92,120,400	0	34
67105	July 10-19	0.36	1.00	1.43	2.50	28	14	7	4	16,90,300,1600	0	10
67107	July 31-Aug. 10	0.18	0.00	4.67	44.00	22	4	3	1	5000	0	44
67109	Aug. 21-30	0.00	0.00	1.33	32.00	16	8	3	1	250,260	0	32
67111	Sept. 11-17	2.563 (0.07)*	0.50	2.00	--	16 (15)*	8	1	0	0,400	0	40
67113	Oct. 2-9	0.17	1.25	13.20	34.00	12	8	5	1	150,880,3500 12 000	0	52
67115	Oct. 24-30	0.18	1.40	14.67	17.00	11	5	3	1	340	0	42
68102	May 17-22	0.11	1.09	12.00	53.50	19	11	2	2		0	54
68108	July 29-Aug. 3	0.00	1.25	6.43	15.00	13	12	7	1		0	24
68112	Nov. 5-10	0.40	7.29	--	72.00	15	7	0	1		0	72
69110	Oct. 15-20	0.13	0.14	1.29	4.00	45	14	7	1	0,2,8,10,48	1	6

* Station with a fecal coliform concentration of 40 has been excluded.

Note: Zones are those determined by inclusion of all stations for which total coliform concentration was measured. See Table 58.

Table 63. Summary of Total Coliform Concentrations at Stations for Which Both Total and Fecal Coliforms Were Measured

	Mean for zone				Number in zone				Zone	Zones 1 to 4	
Cruise	1	2	3	4	1	2	3	4	5	Min.	Max.
67103	1.17	19.92	56.40	160.00	24	13	5	1	270,330,600,8300	0	160
67105	0.14	4.29	15.43	36.25	28	14	7	4	300,500,800,11 000	0	45
67107	0.50	9.50	34.33	100.00	22	4	3	1	6300	0	100
67109	0.44	6.00	36.33	220.00	16	8	3	1	750,22 000	0	220
67111	0.38 (0.40)*	6.88	35.00	--	16 (15)*	8	1	0	900,1000	0	35
67113	0.17	4.50	28.40	180.00	12	8	5	1	500,6000,12 000 51 000	0	180
67115	0.27	4.00	38.00	70.00	11	5	3	1	2600	0	70
68102	0.63	7.46	35.50	79.50	19	11	2	2	79,80	0	80
68108	1.46	14.00	44.29	260.00	13	12	7	1		0	260
68112	1.40	12.86	--	150.00	15	7	0	1		0	150
69110	2.04	23.50	63.57	130.00	45	14	7	1	210,480,520, 540,770	1	130

* Station with a fecal coliform concentration of 40 has been excluded.

Note: Zones are those determined by inclusion of all stations for which total coliform concentration was measured. See Table 58.

Table 64. Ratio of Fecal Coliform Concentration to Total Coliform Concentration

		All stations		Common stations			Ratio	
Cruise	Zone	Mean total (1)	n	Mean total (2)	Mean fecal (3)	n	(3)/(1)	(3)/(2)
67103	1	1.16	113	1.71	0.17	24	0.15	0.10
	2	19.68	37	19.92	1.23	13	0.06	0.06
	3	61.20	10	56.40	7.00	5	0.11	0.12
67105	1	0.21	97	0.14	0.36	28	1.71	2.57
	2	3.81	48	4.29	1.00	14	0.26	0.23
	3	14.86	14	15.43	1.43	7	0.10	0.09
67107	1	0.36	70	0.50	0.18	22	0.50	0.36
	2	6.44	23	9.50	0.00	4	0	0
	3	34.33	3	34.33	4.67	3	0.14	0.14
67109	1	0.52	71	0.44	0.00	16	0	0
	2	6.41	17	6.00	0.00	8	0	0
	3	37.00	6	36.33	1.33	3	0.04	0.04
67111	1	0.33	76	0.40	0.07	15	0.21	0.17*
	2	6.29	17	6.88	0.50	8	0.08	0.07
67113	1	0.18	57	0.17	0.17	12	0.94	1.00
	2	4.22	27	4.50	1.25	8	0.30	0.28
	3	33.33	6	28.40	13.20	5	0.40	0.46
67115	1	0.33	48	0.27	0.18	11	0.55	0.67
	2	6.50	18	4.00	1.40	5	0.22	0.35
	3	40.00	6	38.00	14.67	3	0.37	0.39
68102	1	0.44	43	0.63	0.11	19	0.25	0.17
	2	7.14	14	7.46	1.09	11	0.15	0.15
	3	43.33	3	35.50	12.00	2	0.28	0.34
68108	1	1.52	27	1.46	0.00	13	0	0
	2	13.23	22	14.00	1.25	12	0.09	0.09
	3	43.42	12	44.29	6.43	7	0.15	0.15
68112	1	0.98	40	1.40	0.40	15	0.41	0.29
	2	14.67	9	12.86	7.29	7	0.50	0.57
	3	44.00	2	-	-	0	-	-
69110	1	3.02	46	3.07	1.02	45	0.34	0.33
	2	23.50	14	23.50	1.00	14	0.04	0.04
	3	63.57	7	63.57	2.00	7	0.03	0.03

* Station with a fecal coliform concentration of 40 has been excluded.

Table 65. Turbidity and Total Coliforms by Zone Using Only Stations at Which Both Turbidity and Total Coliforms Were Measured

		Mean for zone				Number in zone				Zones 1 to 4	
Variable	Cruise	1	2	3	4	1	2	3	4	Min.	Max.
Turbidity	68102	1.20	2.04	2.27	6.00	43	12	3	2	0.1	9.2
	68108	0.33	0.46	0.51	0.20	27	22	11	1	0.1	1.8
	68109	0.65	0.65	-	-	23	5	0	0	0.0	3.0
	68112	1.74	3.29	2.35	12.00	40	9	2	1	0.2	12.0
Total coliforms	68102	0.44	7.17	43.33	79.50	43	12	3	2	0	80
	68108	1.52	13.23	44.82	260.00	21	22	11	1	0	260
	68109	2.83	48.50	-	-	23	4	0	0	0	110
	68112	0.98	14.67	44.00	150.00	40	9	2	1	0	150

Total number of common stations are 60, 61, 27 and 52 for cruises 102, 108, 109 and 112 compared with 62, 62, 28 and 52 for total coliforms.

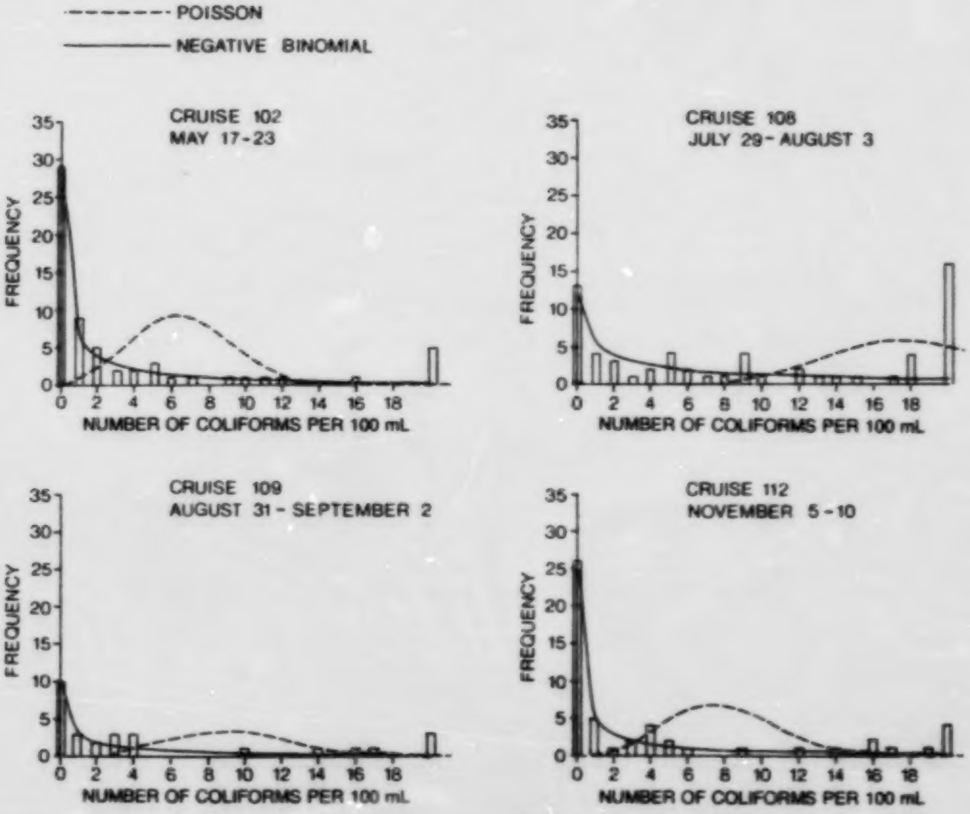


Figure 63. The observed frequency distribution and the fitted Poisson and negative binomial frequency distributions, 1968 total coliform data.

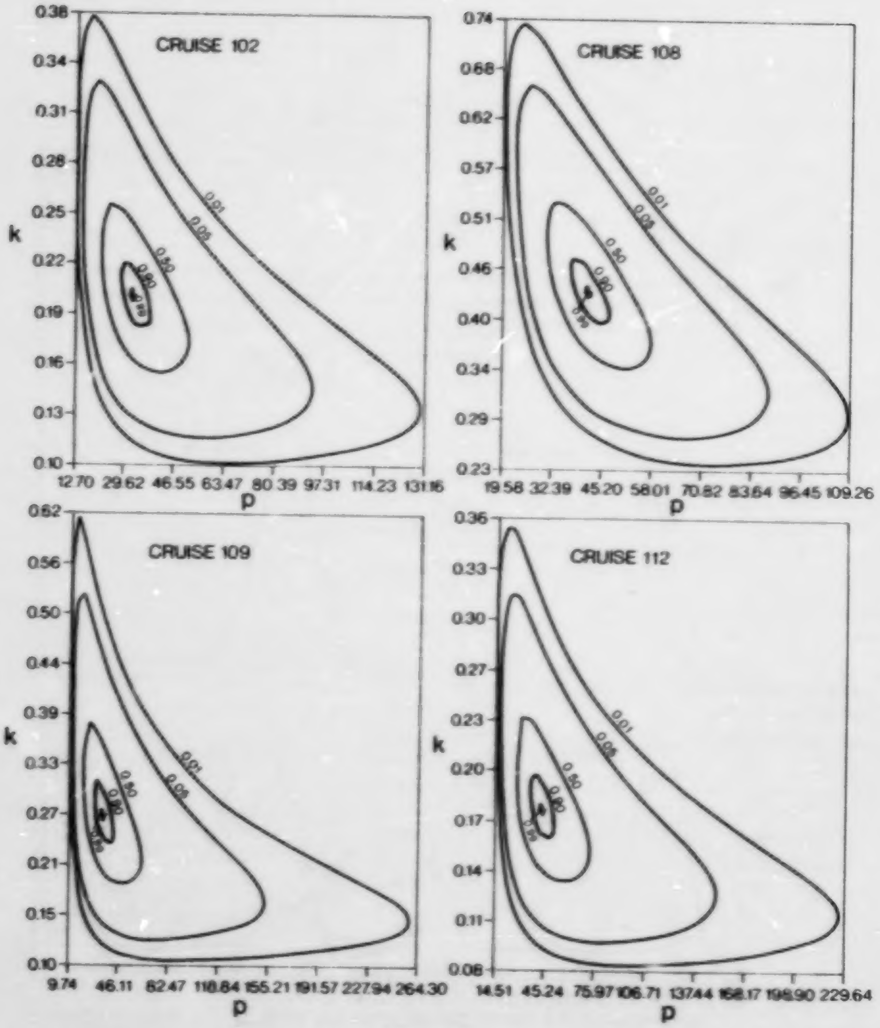


Figure 64. Contours of constant relative likelihood R(p,k) = c, c = 0.01, 0.05, 0.50, 0.90 and 0.99, for p,k parameters of the negative binomial distribution and total coliform data for cruises 102, 108, 109 and 112 of 1968.

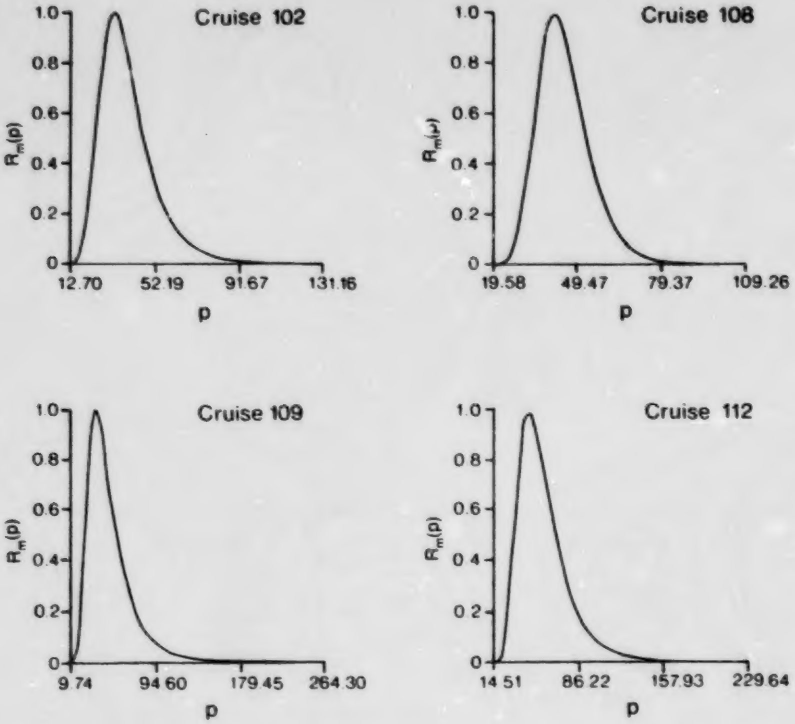


Figure 65a. Relative maximum likelihood functions of p, $R_m(p)$, for p a parameter of the negative binomial distribution and total coliform data for cruises 102, 108, 109 and 112 of 1968.

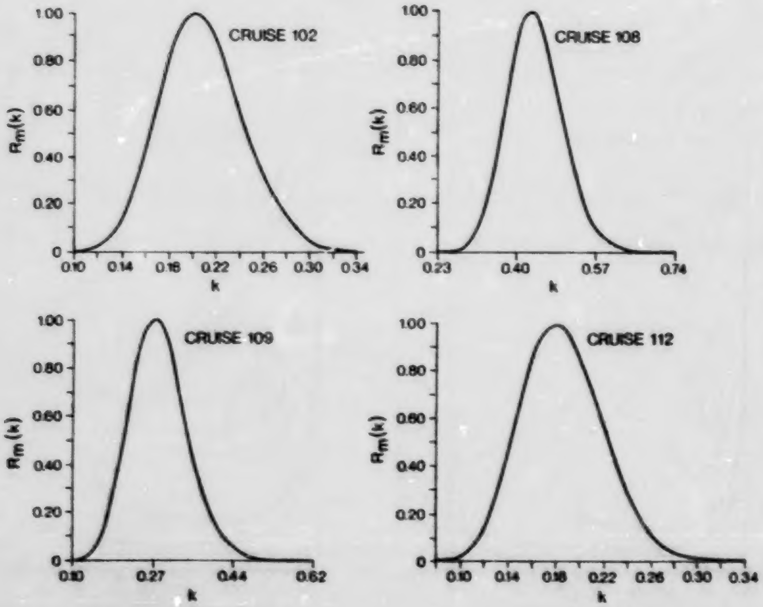
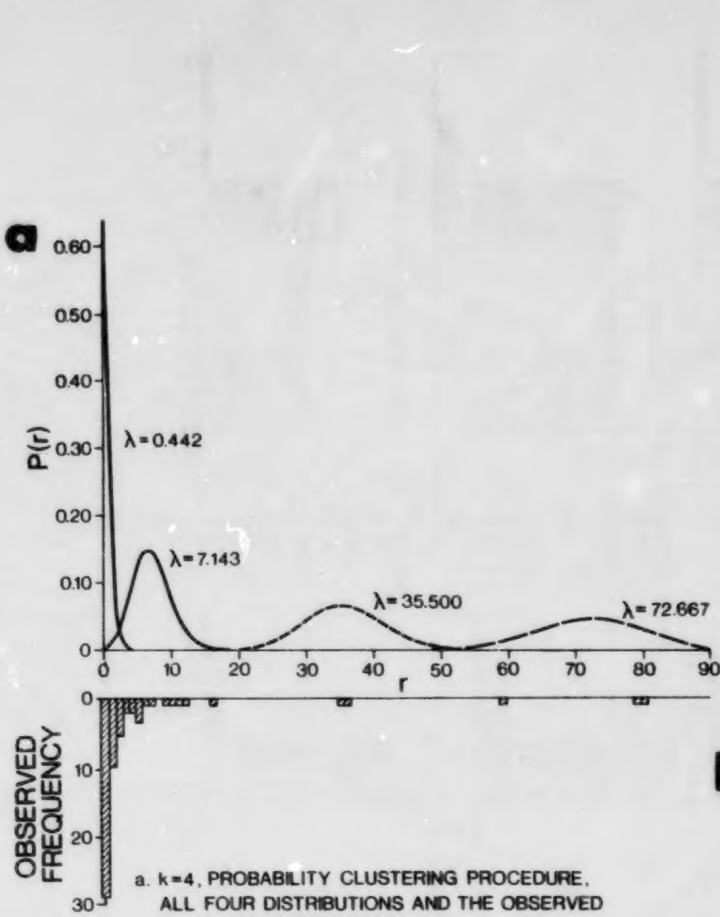
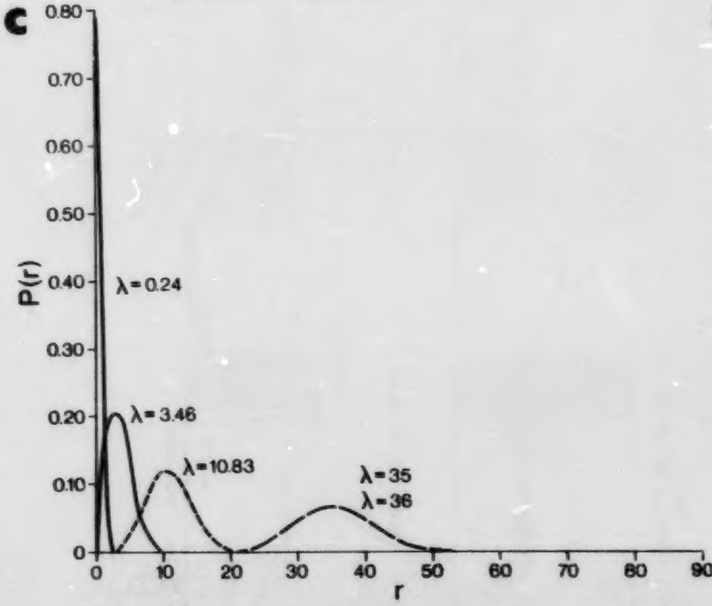


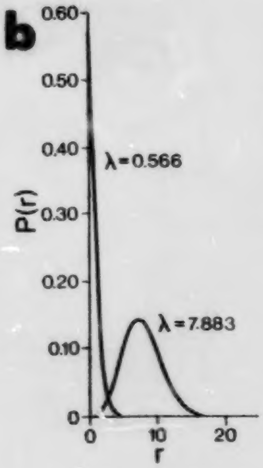
Figure 65b. Relative maximum likelihood functions for k, $R_m(k)$, for k a parameter of the negative binomial distribution and total coliform data for cruises 102, 108, 109 and 112 of 1968.



a. k=4, PROBABILITY CLUSTERING PROCEDURE, ALL FOUR DISTRIBUTIONS AND THE OBSERVED FREQUENCY DISTRIBUTION



c. k=7, PROBABILITY CLUSTERING PROCEDURE, FIRST 5 DISTRIBUTIONS



b. k=4, GROUPING BY D^2 CRITERION, FIRST 2 DISTRIBUTIONS

Figure 66. Probability distributions corresponding to groups determined for cruise 102, 1968, where P(r) is the probability assuming a Poisson distribution with mean λ shown adjacent to curve.

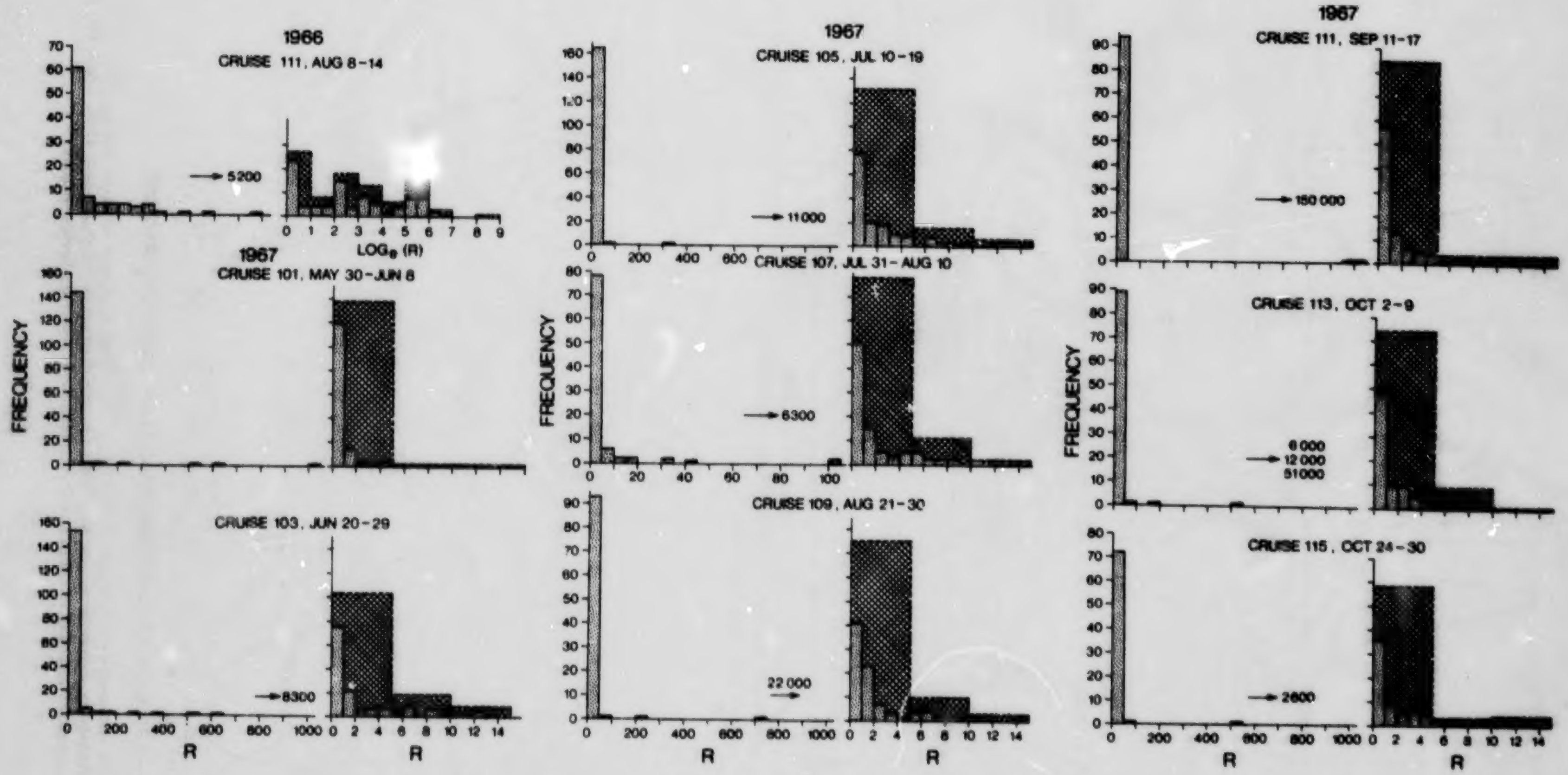


Figure 67. Observed frequency distributions for total coliform concentration (R) or log$_e$R shown on two scales.

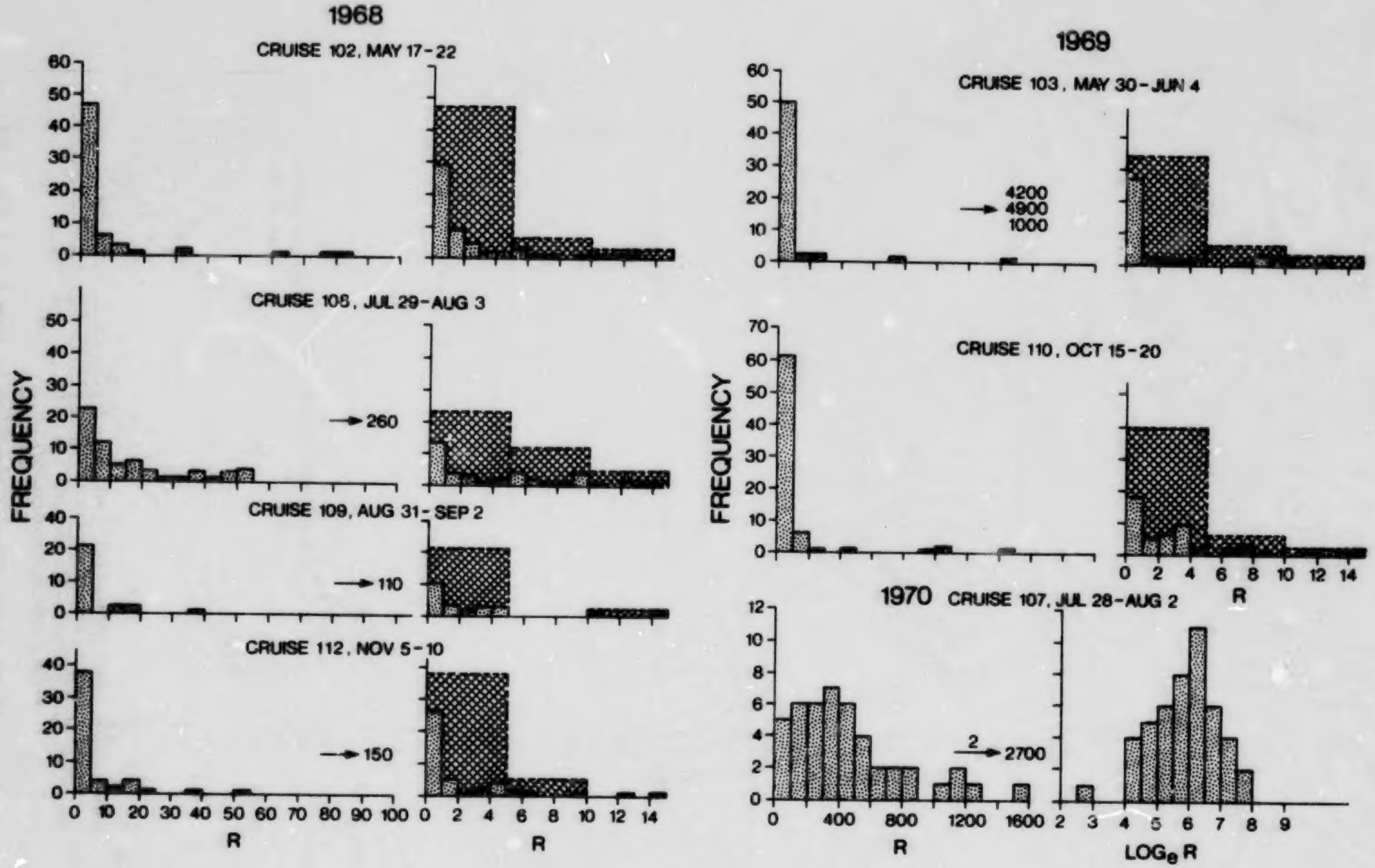


Figure 37. Continued

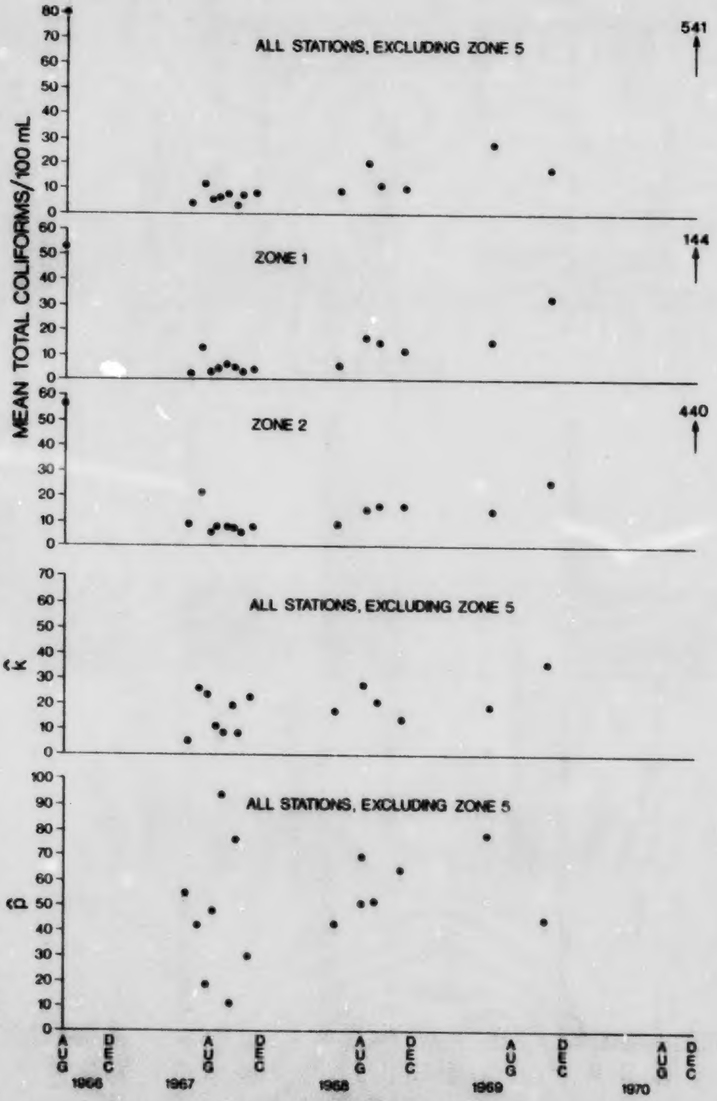
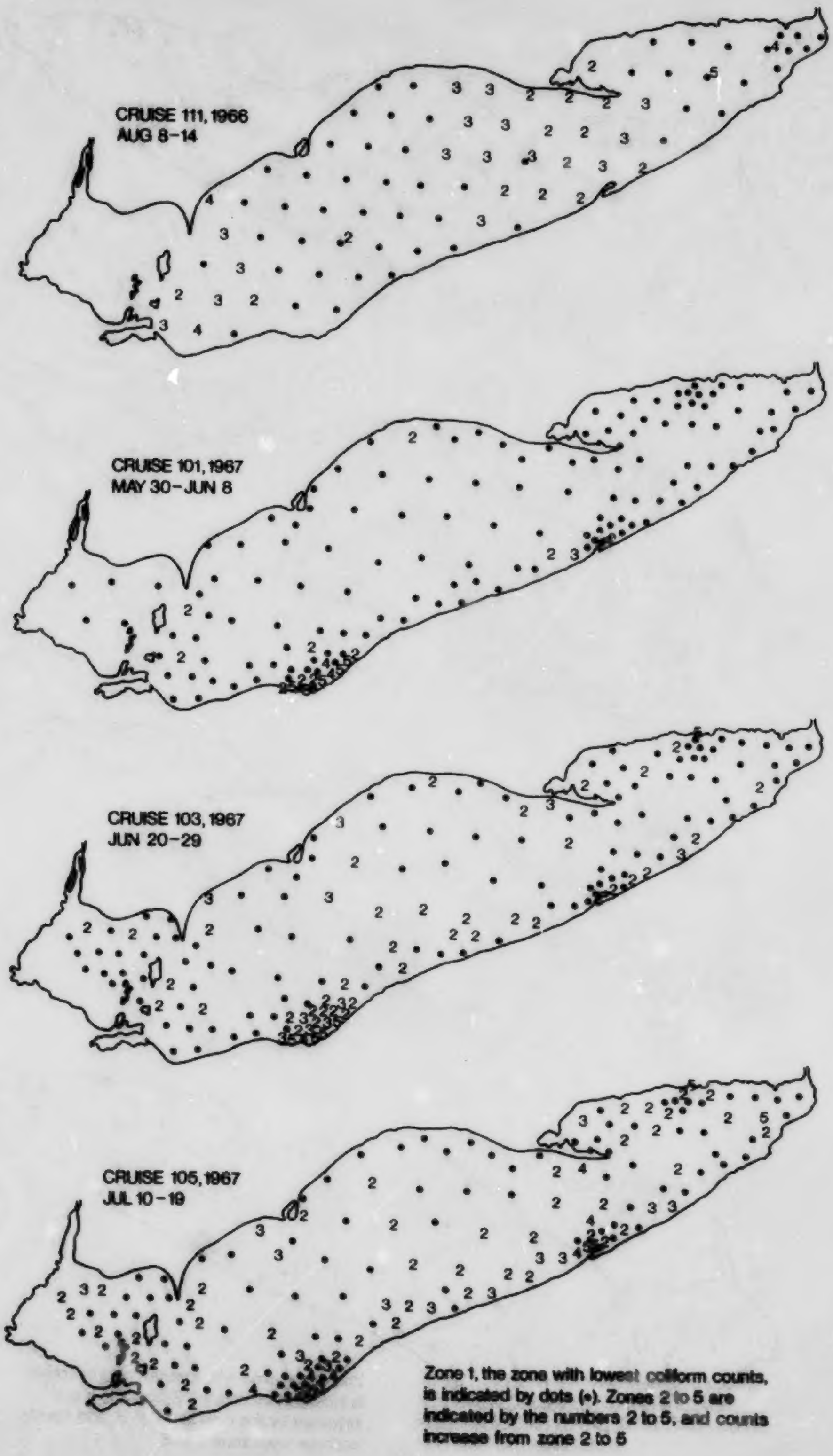


Figure 68. Mean total coliform concentrations for zones 1 to 4, zone 1 and zone 2, and maximum likelihood estimates of the parameters of the negative binomial distribution by cruise.



CRUISE 111, 1966
AUG 8–14

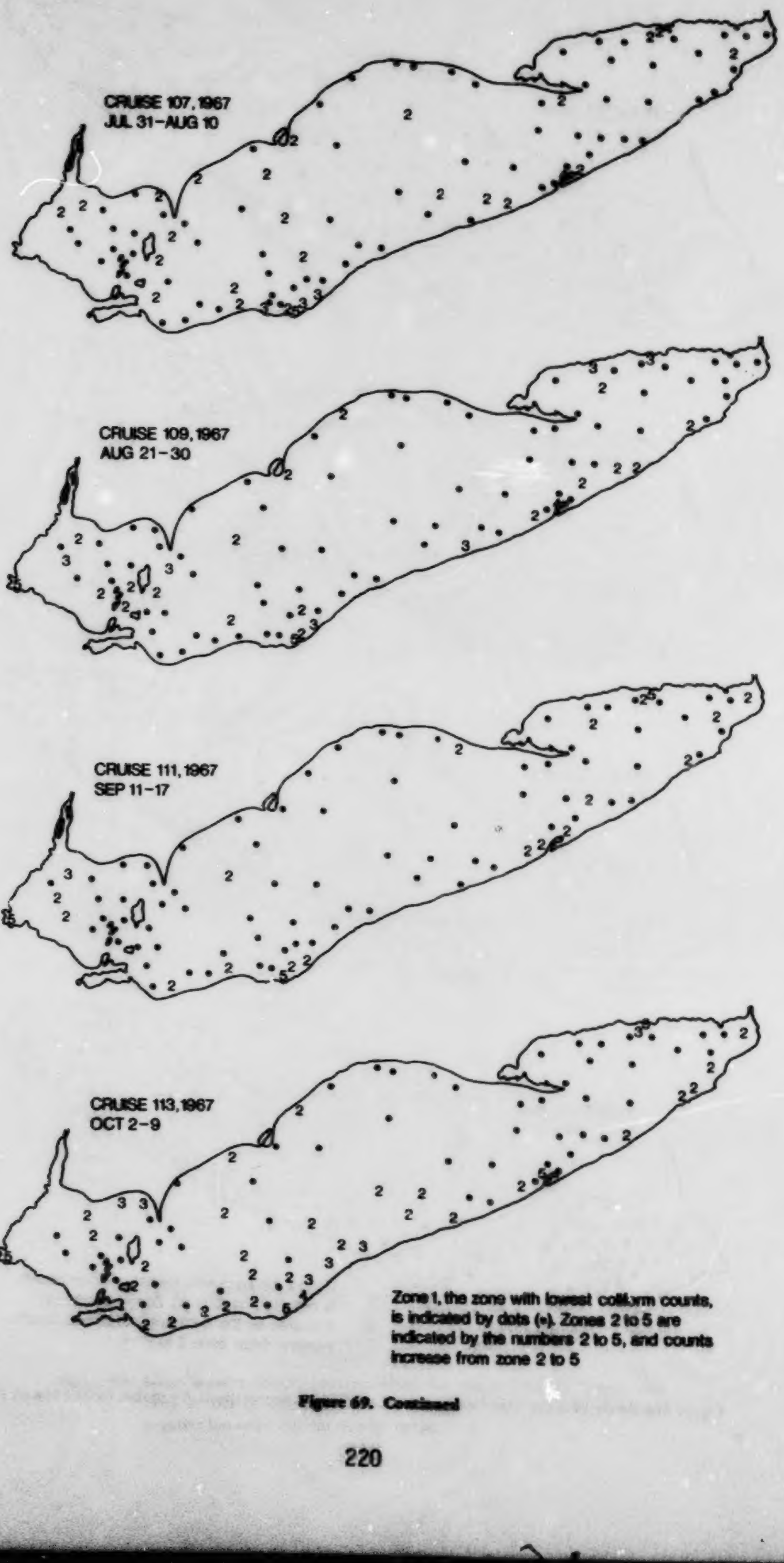
CRUISE 101, 1967
MAY 30–JUN 8

CRUISE 103, 1967
JUN 20–29

CRUISE 105, 1967
JUL 10–19

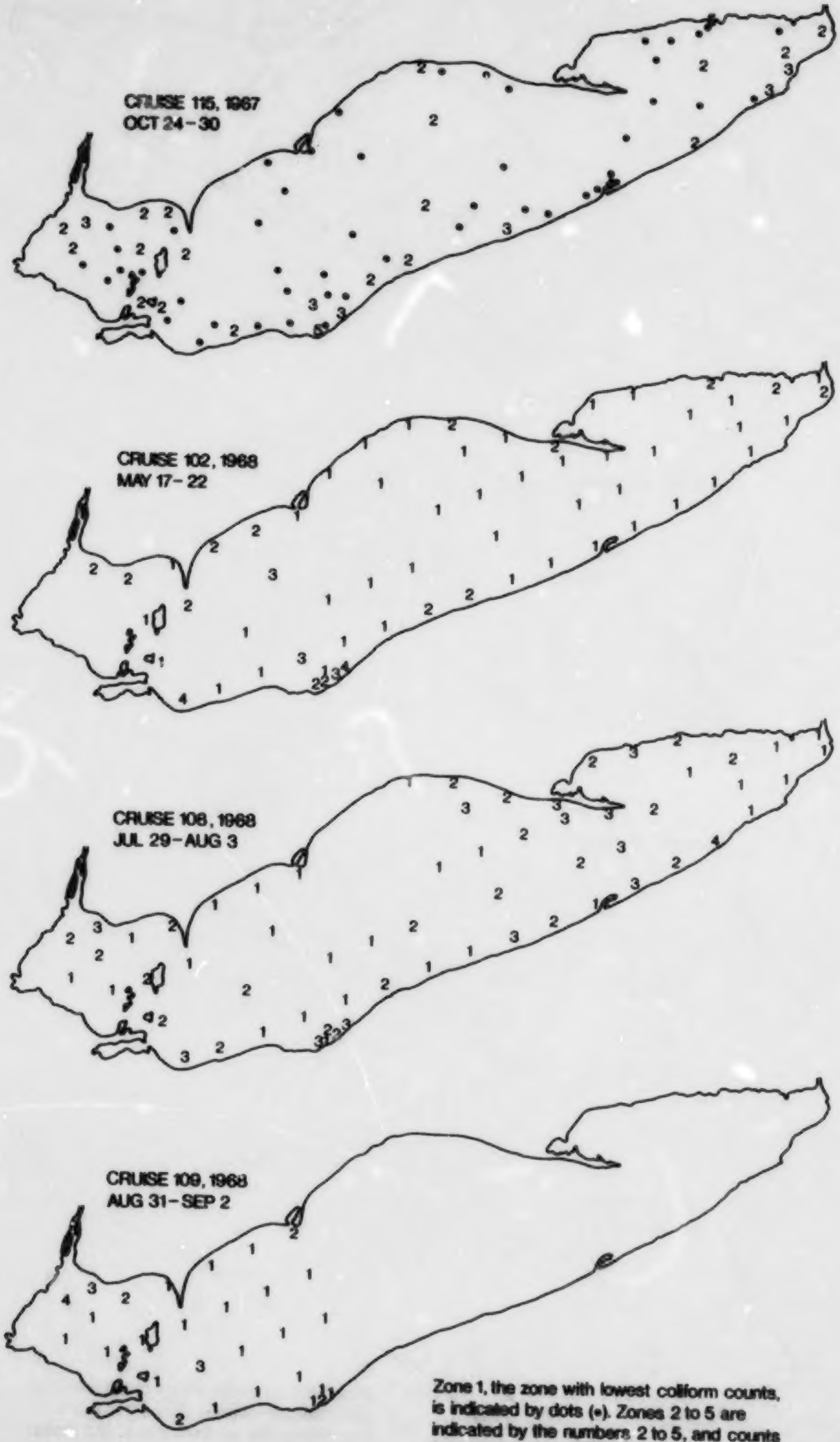
Zone 1, the zone with lowest coliform counts, is indicated by dots (•). Zones 2 to 5 are indicated by the numbers 2 to 5, and counts increase from zone 2 to 5

Figure 69. Zones of Lake Erie based on total coliform concentrations, by cruise, from 1966 to 1970.


CRUISE 107, 1967
JUL 31–AUG 10
CRUISE 109, 1967
AUG 21–30
CRUISE 111, 1967
SEP 11–17
CRUISE 113, 1967
OCT 2–9

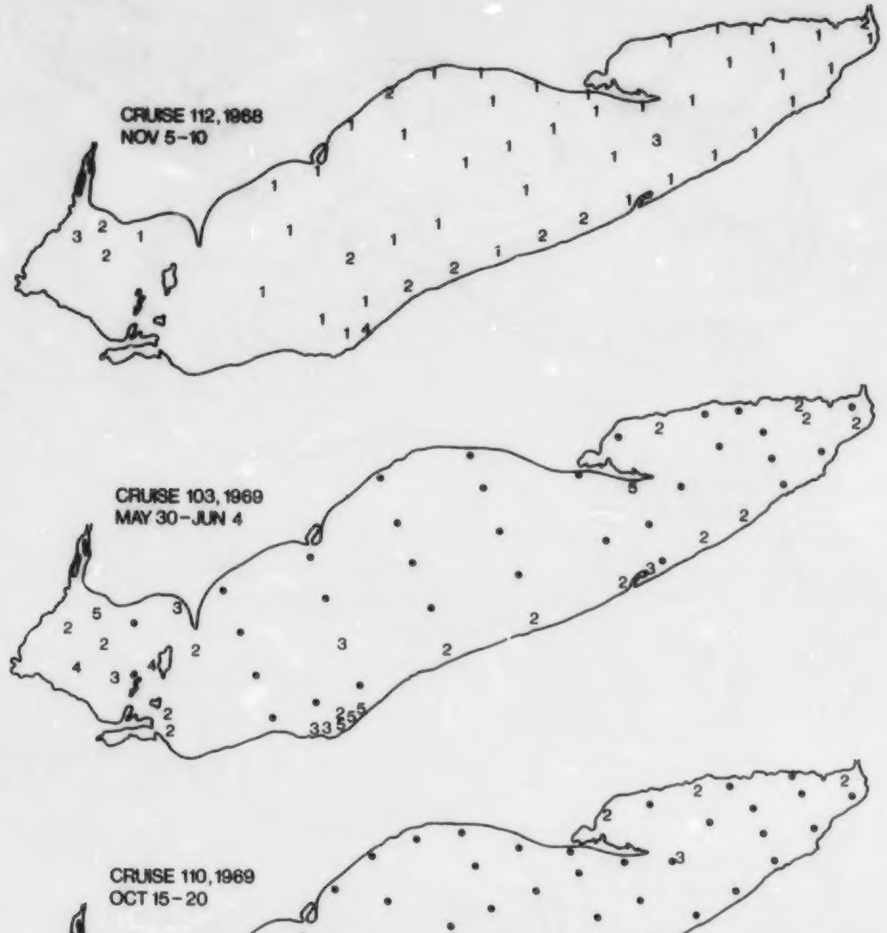
Zone 1, the zone with lowest coliform counts, is indicated by dots (•). Zones 2 to 5 are indicated by the numbers 2 to 5, and counts increase from zone 2 to 5

Figure 69. Continued


CRUISE 115, 1967
OCT 24-30
CRUISE 102, 1968
MAY 17-22
CRUISE 108, 1968
JUL 29-AUG 3
CRUISE 109, 1968
AUG 31-SEP 2

Zone 1, the zone with lowest coliform counts, is indicated by dots (•). Zones 2 to 5 are indicated by the numbers 2 to 5, and counts increase from zone 2 to 5

Figure 69. Continued


CRUISE 112, 1968
NOV 5-10
CRUISE 103, 1969
MAY 30-JUN 4
CRUISE 110, 1969
OCT 15-20
CRUISE 107, 1970
JUL 28-AUG 2

Zone 1, the zone with lowest coliform counts, is indicated by dots (•). Zones 2 to 5 are indicated by the numbers 2 to 5, and counts increase from zone 2 to 5

Figure 69. Continued

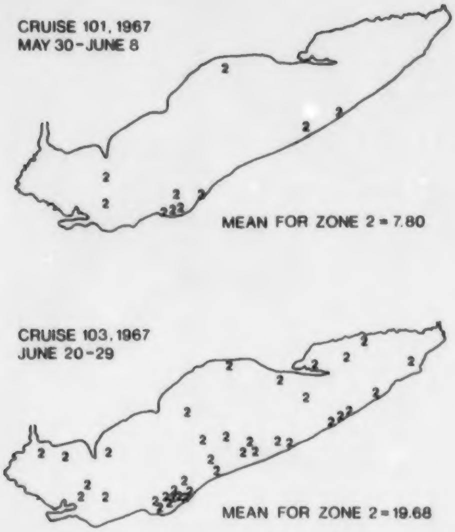

CRUISE 101, 1967
MAY 30-JUNE 8
MEAN FOR ZONE 2 = 7.80
CRUISE 103, 1967
JUNE 20-29
MEAN FOR ZONE 2 = 19.68

Figure 70. Examples of zone 2 characteristic of the spring and peak total coliform distributions.

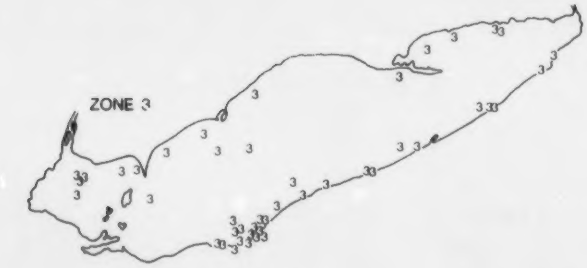

ZONE 3
ZONE 4
ZONE 5
5 TIMES
5 TIMES
8 TIMES

Figure 71. Zones 3, 4 and 5 based on the total coliform concentrations plotted together for all 1967 cruises.

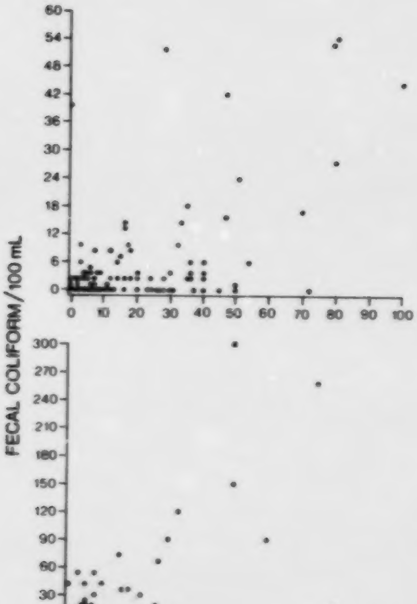


Figure 72. Fecal against total coliform plotted on two ranges of total coliform concentrations, 1967 and 1968 cruises.

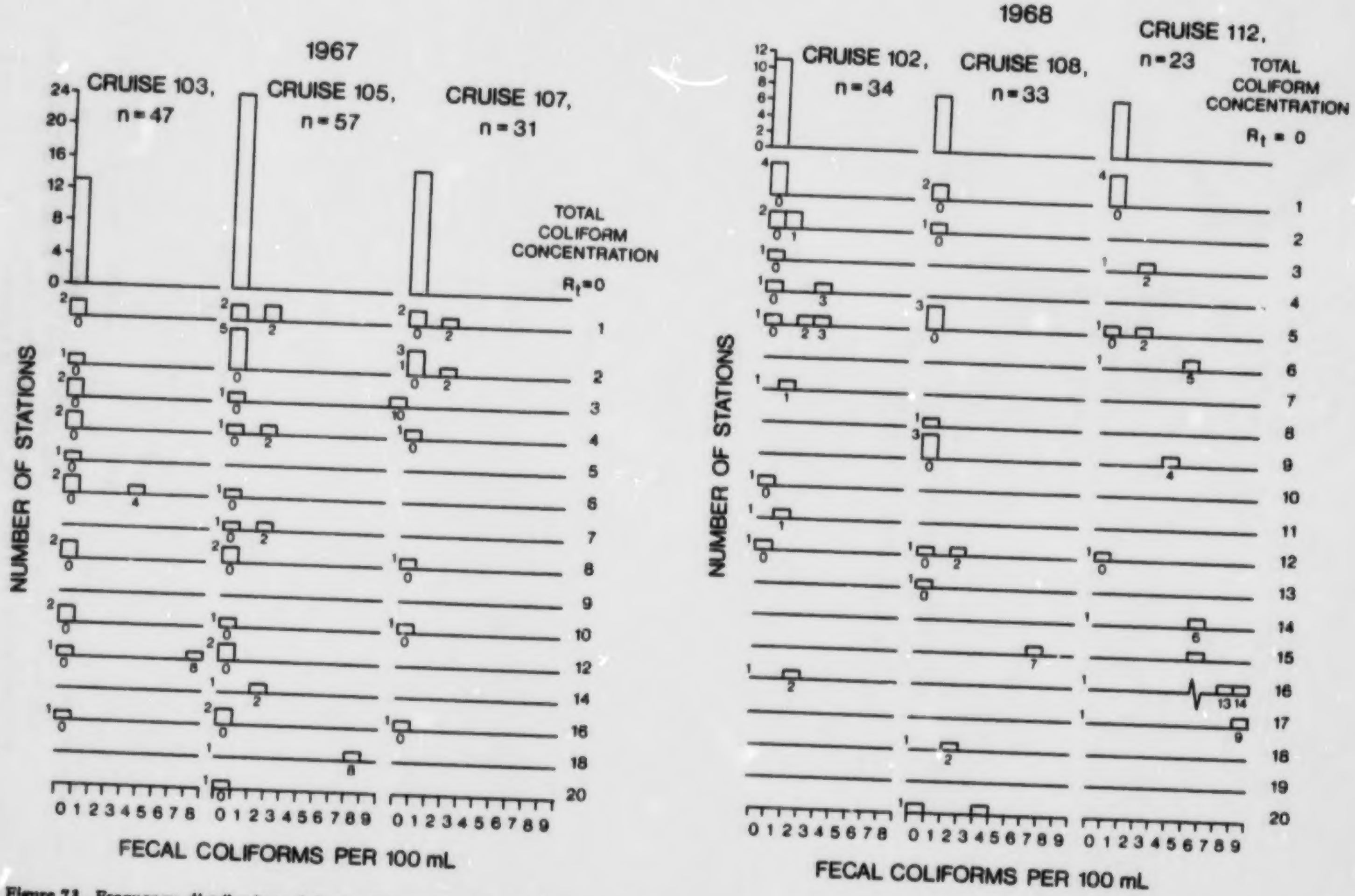


Figure 73. Frequency distribution of fecal coliforms conditional on the total coliform concentration (R_t) shown for $R_t < 20$ and for three 1967 cruises and three 1968 cruises (n = total number of common stations).

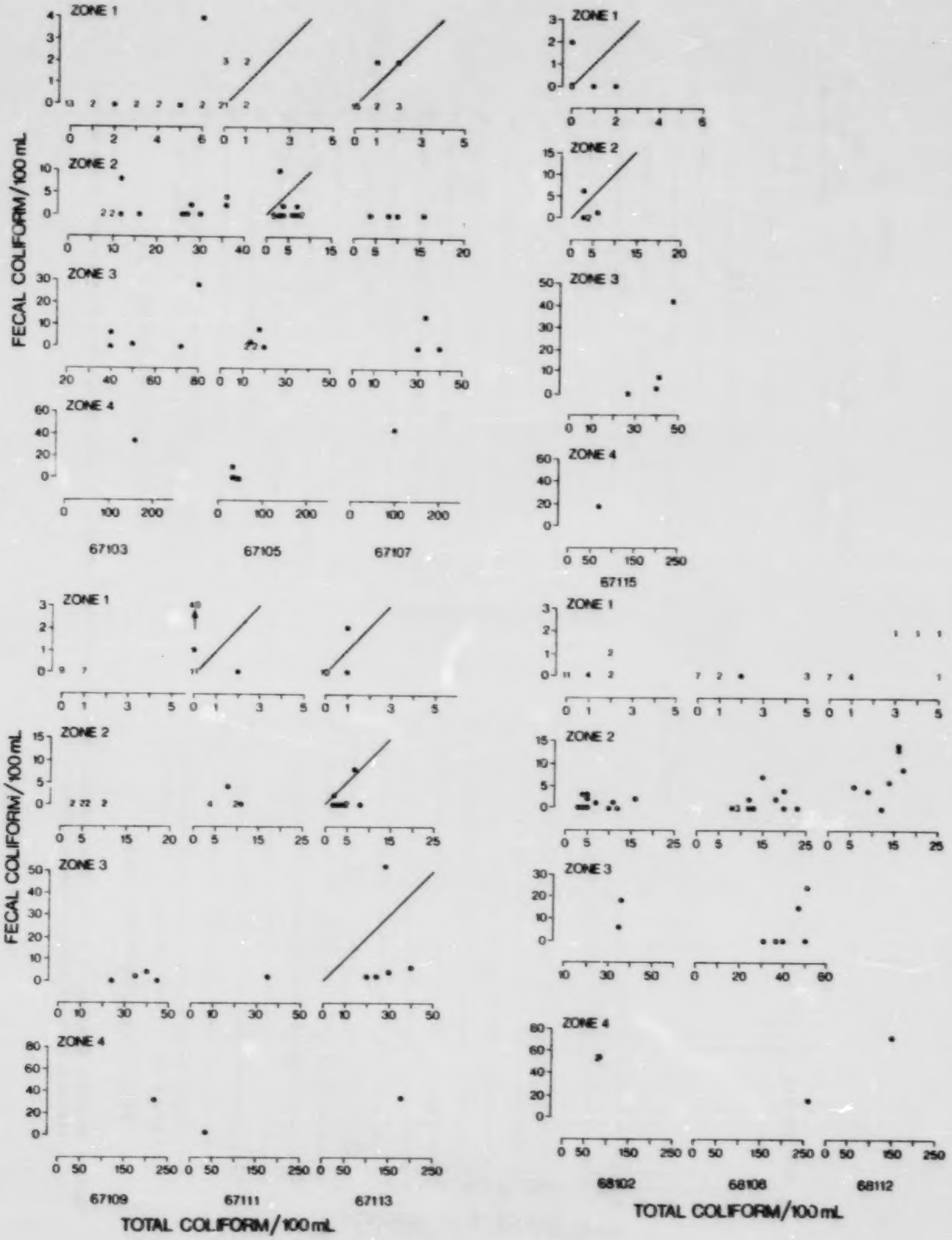


Figure 74. Fecal coliform concentrations plotted against total coliform concentrations by zone and cruise for 1967 and 1968.

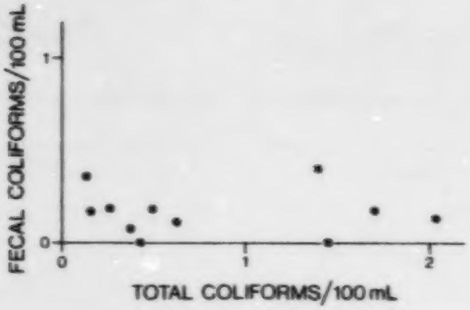


Figure 75. Mean fecal coliform concentration against mean total coliform concentration for zone 1 of the 1967–1969 cruises.

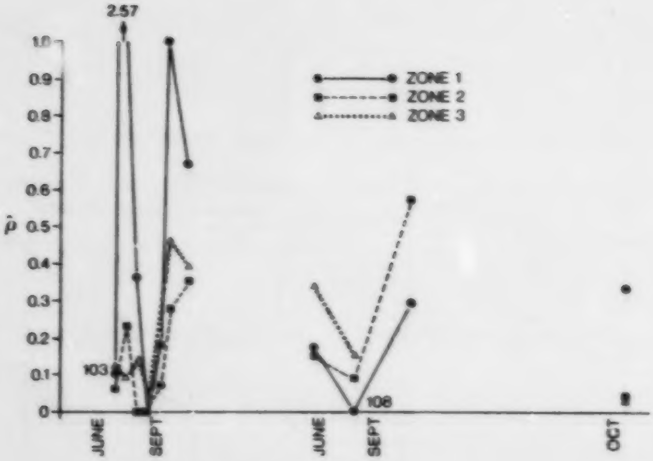


Figure 76. Ratio of fecal coliform concentration to total coliform concentration plotted against cruise date for zones 1, 2 and 3 using data from common stations.

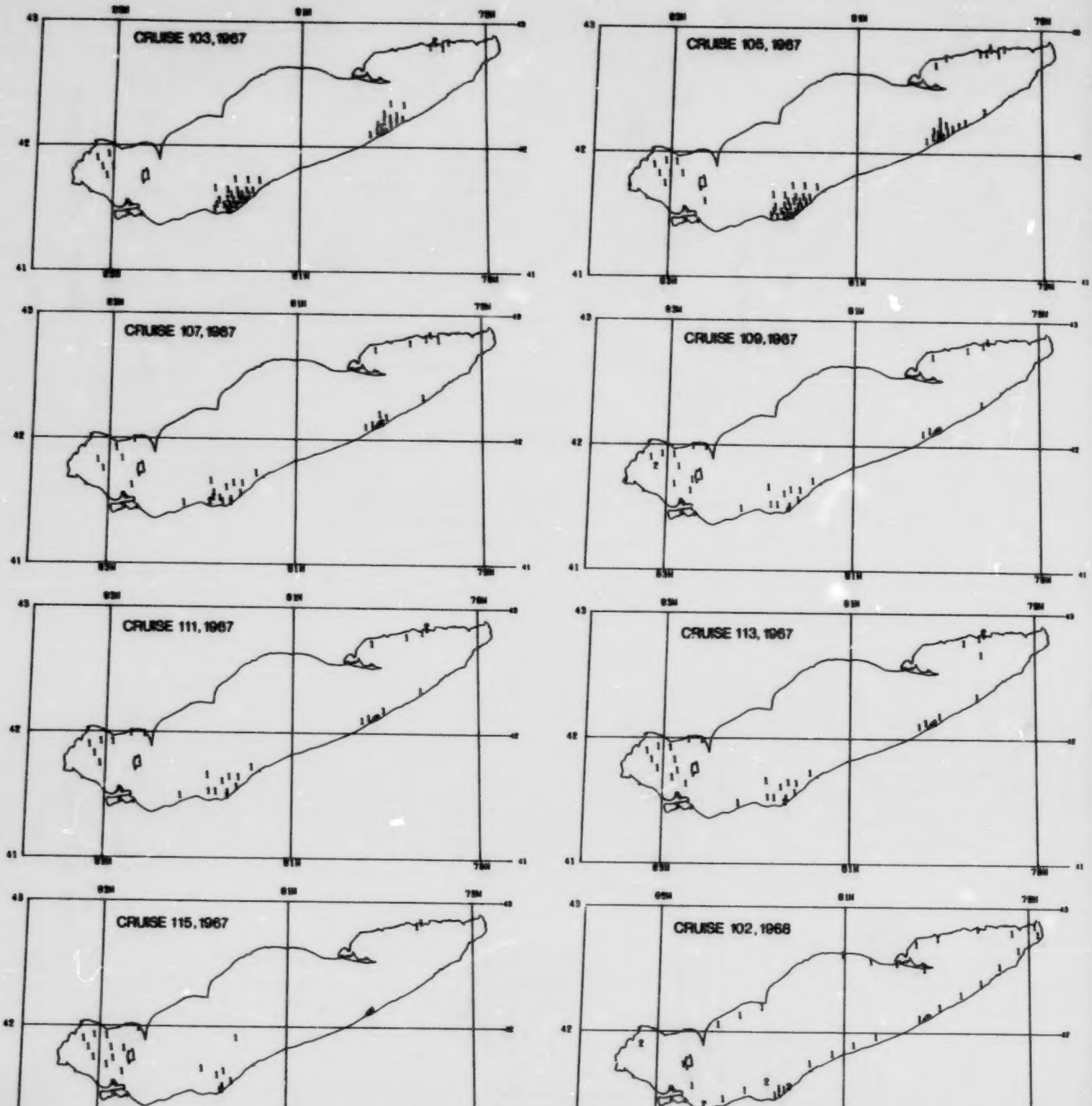
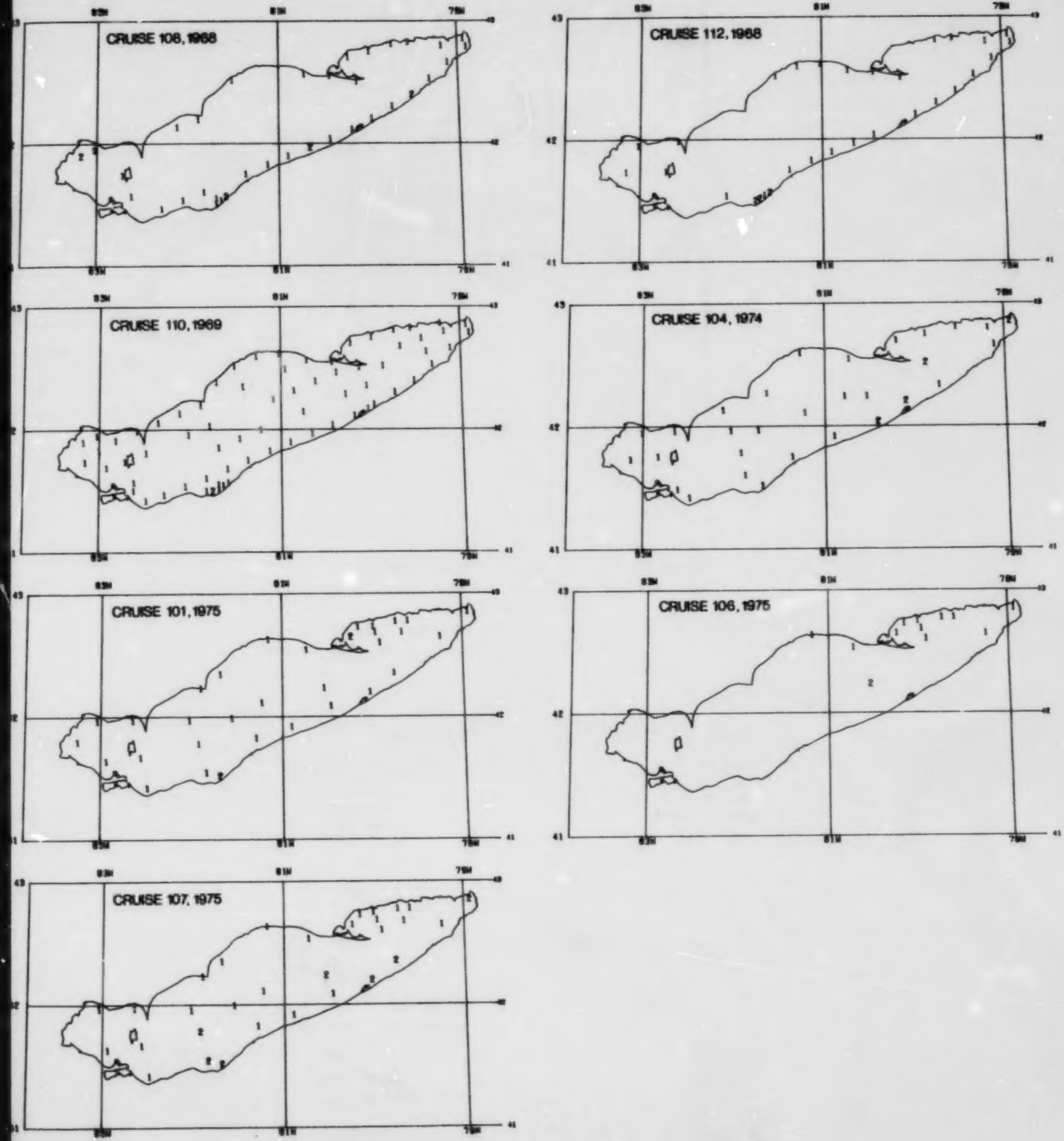

CRUISE 103, 1967
CRUISE 105, 1967
CRUISE 107, 1967
CRUISE 109, 1967
CRUISE 111, 1967
CRUISE 113, 1967
CRUISE 115, 1967
CRUISE 102, 1968
83W
81W
79W
43
42
41

Figure 77. Location of stations at which fecal coliform concentrations were determined.



CRUISE 108, 1968

CRUISE 112, 1968

CRUISE 110, 1969

CRUISE 104, 1974

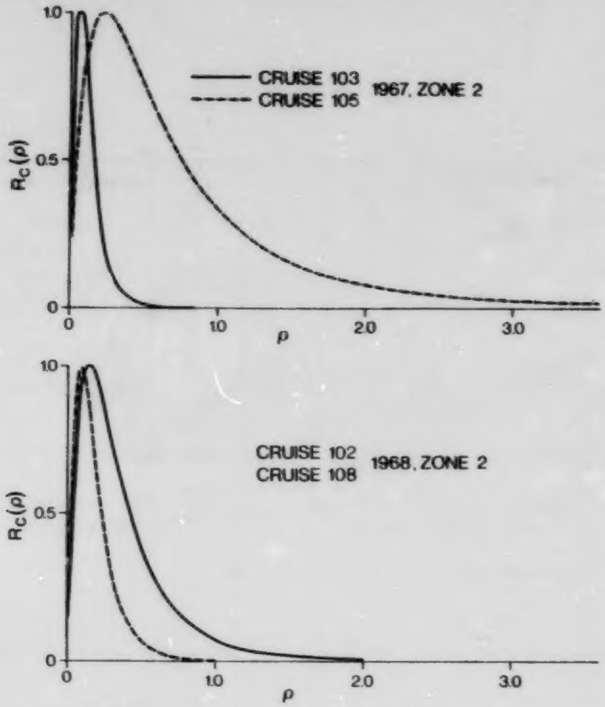
CRUISE 101, 1975

CRUISE 106, 1975

CRUISE 107, 1975

83W
81W
79W
43
42
41

Figure 77. Continued



$R_c(\rho)$ >0.5	for ρ
— 103,1967	0.02 to 0.16
—— 105,1967	0.05 to 0.75
—— 102,1968	0.04 to 0.43
— 108,1968	0.03 to 0.24

Figure 78. Relative conditional likelihood function, $R_c(\rho)$, for the ratio of fecal to total coliform concentration for four cruises.

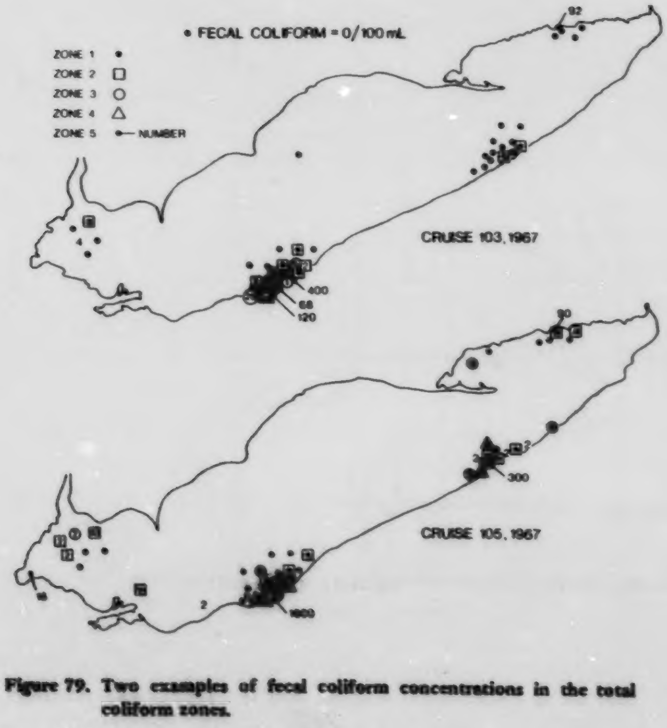

• FECAL COLIFORM = 0/100 mL
ZONE 1 •
ZONE 2 □
ZONE 3 ○
ZONE 4 △
ZONE 5 •— NUMBER
92
CRUISE 103, 1967
400
68
120
90
2
300
CRUISE 105, 1967
98
2
1960

Figure 79. Two examples of fecal coliform concentrations in the total coliform zones.

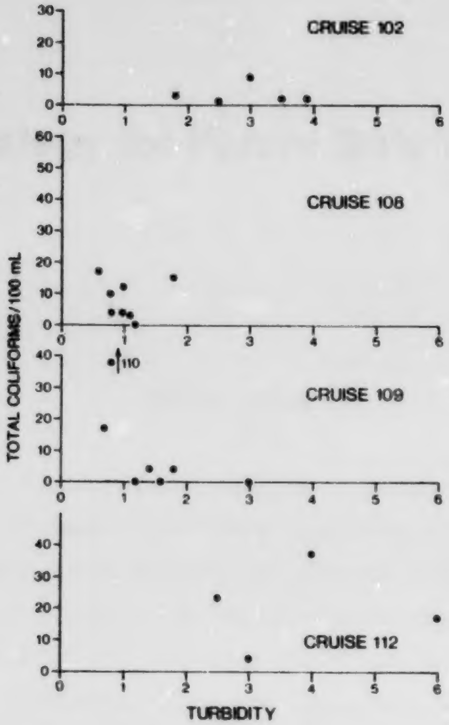


Figure 80. Total coliform concentration against turbidity for stations in the Western Basin, 1968 cruises.

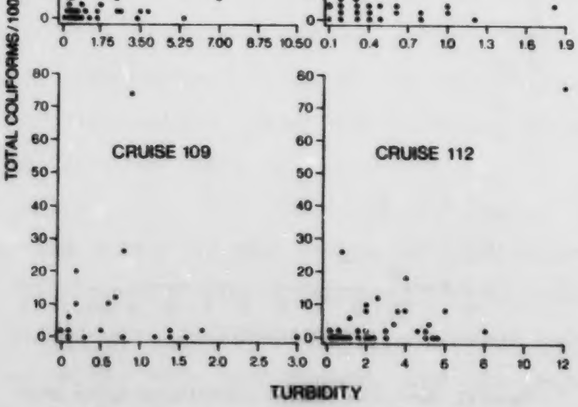


Figure 81. Total coliform concentration against turbidity for 1968 cruises.

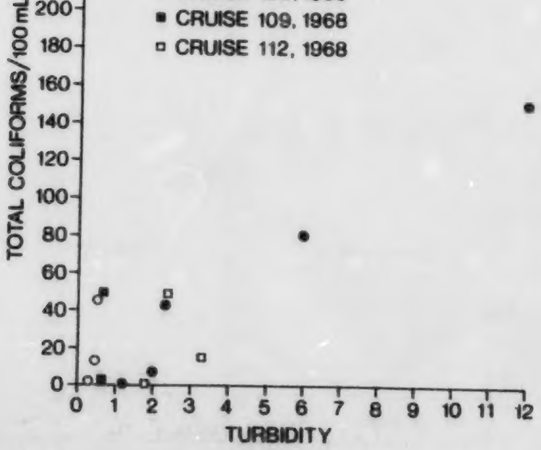


Figure 82. Mean total coliform concentration against mean turbidity for zones 1 to 4 of the 1968 cruises using common stations in the zones determined from all the total coliform data.

CHAPTER 7

Sampling Strategy for Future Data Collection

by A.H. El-Shaarawi

INTRODUCTION

The water quality of a lake or a large body of water can only be estimated by sampling. Many agencies concerned with water quality such as the Canada Centre for Inland Waters conduct several sampling surveys each year. A survey starts by determining a sampling strategy which specifies (i) the parameters to be measured, (ii) the number of sampling stations, (iii) the locations of the stations, and (iv) for each station, the depths at which water samples are to be collected.

The accuracy of the results obtained by the survey is influenced by several factors. The first, and probably the most important, is the choice of the number of stations and of their locations. This is particularly true if the limnological variables show a very high degree of spatial variability over the lake. The inappropriate choice of the locations of the stations will produce data which will not give an adequate description of the condition of the lake and probably will produce a biased picture and, hence, lead to incorrect results. Therefore, it is of the utmost importance to use all the available information about the spatial variability of the lake in designing future surveys.

The second factor is the length of time required to conduct the survey. This factor is important, since the limnological properties of water are influenced by a number of environmental variables such as wind, rain and runoff, and the majority of these variables cannot be controlled or predicted, especially for a long period of time. Hence, it is appropriate to minimize the effects of these variables by shortening the time needed to complete the survey. It is evidently clear that the effect of this factor is in conflict with that of the first factor, i.e., the larger the number of stations and the longer the distances between them, the longer the time required for completing the survey.

The accuracy of the survey is influenced by many other important factors which can be summarized as the care and consistency with which the water samples are collected, the sample analysis is performed, and the results are recorded. The effects of these factors will not be considered in this chapter.

To develop a strategy for sampling, the objectives of the survey should be precisely specified. These objectives and the monetary and practical constraints can be incorporated into a mathematical expression which specifies the number of stations that need to be sampled, and their locations in the lake.

In this chapter, the question is investigated concerning how to use past information about the spatial and temporal variabilities of limnological data to plan a strategy for future data collection when the aim is to estimate the areal weighted mean value of a single limnological variable. The results presented here are intended simply as a preliminary discussion of the problem of survey design with emphasis on some general principles. Two types of representation of the spatial pattern are assumed: the first regards the lake as a set of discrete "homogeneous" zones and the second assumes that the values taken by the limnological variable can be regarded as a realization of a continuous random variable with a mean which can be expressed as a continuous function (surface) of the position of the sampling stations. The multivariate generalization of this problem is very difficult, requiring joint statistical and limnological research and a better data base than that available for Lake Erie. Hence, this case is not considered here.

The statistical approaches are illustrated and discussed using data on coliform counts, temperature and chlorophyll a. The next section presents a hypothetical example to illustrate the principles of sampling design. To be specific, it is assumed that the interest is the estimation of the mean bacteria of a lake; the discussion, however, is valid for any other variable.

EXAMPLE TO ILLUSTRATE THE PRINCIPLES OF SAMPLING

Suppose that the objective of the survey is to obtain an estimate of the number of bacteria, μ, per unit volume. The usual procedures are to collect a number of water samples and then to estimate μ by calculating the arithmetic or the geometric mean of the data. It is important to remember that the adequacy of such an averaging process depends upon where the water samples were collected. To illustrate this point, consider as an example an idealized lake shaped like a box of length ℓ, width W and depth D, with uniform vertical distribution of bacteria. Assume further that the number of bacteria per unit volume, the density, follow the linear relationship:

$$\mu_\ell = \alpha + \beta_\ell \qquad 0 \leqslant \ell \leqslant L \tag{7.1}$$

where μ_ℓ is the density of bacteria at a point ℓ units from the end of the lake taken as the origin for all distances and α and β are unknown constants. Integrating 7.1 over the lake and dividing by the lake volume gives

$$\mu = \alpha + \beta L/2 \tag{7.2}$$

Suppose that n samples were taken from this lake at distances ℓ_1, $\ell_2 \ldots \ell_n$, and assume further that the random errors are disregarded completely. Under these conditions this sampling strategy gives μ_{ℓ_1}, $\mu_{\ell_2} \ldots \mu_{\ell_n}$ as the observed bacterial counts. When the mean $\bar{\mu} = (\mu_{\ell_1} + \mu_{\ell_2} + \ldots + \mu_{\ell n})/n$ is used to estimate μ, using Equation 7.1 the bias is

$$B = \mu - \bar{\mu} = \beta(\frac{L}{2} - \bar{\ell})$$

The value of B represents the bias which occurs in estimating μ as the result of using this particular sampling design. The bias is zero when $\beta = 0$, which implies that the lake is completely homogeneous or the sampling design is constructed such that $\bar{\ell} = L/2$. The maximum positive bias occurs when $\ell_1 = \ell_2 = \ldots = \ell_n = 0$ and is $\beta L/2$, whereas the maximum negative bias occurs when $\ell_1 = \ell_2 = \ldots = \ell_n = L$ and is $-\beta L/2$. This shows that the locations of the sampling stations play a very important role in obtaining an unbiased estimate for μ when the lake has a very high degree of heterogeneity.

The condition $\bar{\ell} = L/2$ is not sufficient to obtain a good sampling design, as it is possible to take all the water samples at the point $L/2$ to produce an unbiased estimate for μ. Obviously, there is a need for another criterion, which can be obtained by assuming that the values μ_ℓ are not observed exactly, but are subject to random errors with mean 0 and variance σ^2. To be more specific, suppose that it is not possible to observe μ_ℓ but C_ℓ is observed, where $C_\ell = \mu_\ell + \varepsilon_\ell$ and ε_ℓ is a random variable with a mean 0 and variance σ^2. Under these assumptions Equation 7.1 can be rewritten as

$$C_\ell = \alpha + \beta\ell + \varepsilon_\ell \tag{7.3}$$

Given $C_{\ell_1}, C_{\ell_2} \ldots C_{\ell_n}$, the values α and β can be estimated, for example, by regression. Let $\hat{\alpha}$ and $\hat{\beta}$ be the estimates of α and β. The value of μ is then estimated by $\hat{\mu}$, where $\hat{\mu} = \hat{\alpha} + \hat{\beta}L/2$. It can be easily shown that the variance of $\hat{\mu}$ (where α and β are estimated by regression) is

$$\mathrm{var}\ (\hat{\mu}) = \{\frac{\sigma^2}{n} 1 + \frac{(\frac{L}{2} - \bar{\ell})^2}{s^2_\ell}\} \tag{7.4}$$

where $s^2_\ell = \frac{1}{n} \sum_{i=1}^{n} (\ell_i - \bar{\ell})^2$. Equation 7.4 gives the precision of estimating $\hat{\mu}$. The larger the value of var (μ), the smaller the precision; then it is clear that the maximum precision is obtained when $\frac{L}{2} - \bar{\ell} = 0$ and $s^2_\ell \neq 0$. In fact, to maximize the precision, it is necessary to choose the design which makes s^2_ℓ as large as possible and $\frac{L}{2} - \bar{\ell}$ as small as possible. This is achieved by taking only two values for ℓ_i, namely 0 and L, and allocating half of the sample to each end of the lake.

This example shows that an appropriate sampling strategy can be chosen for estimating the bacterial density of a lake when some information is available on the spatial variability of bacterial counts, a subject which is considered in the next section.

SAMPLING STRATEGY WHEN THE LAKE IS DIVIDED INTO ZONES

Suppose that using historical data the lake is divided into κ homogeneous zones by, for example, the method in El-Shaarawi _et al_. (1981). Assume further that the objective of the survey is to estimate the areal weighted bacterial density of the lake. Under the assumption of homogeneity the number of bacteria r_{ji} in the jth sample ($j = 1,2\ldots n_i$) from the ith zone ($i = 1,2\ldots\kappa$) has the Poisson distribution

$$P(r_{ji}) = e^{-\mu_i}\mu_i^{r_{ji}}/r_{ji}$$

with mean μ_i. The parameter μ_i represents the true but unknown bacterial density in the ith zone and is estimated by $\bar{r}_i$ where $\bar{r}_i$ is the arithmetic mean of $r_{1i}, r_{2i}\ldots r_{ni}$. Let n be the number of stations to be sampled during the survey. Then the areal mean density is

$$\bar{r}_\omega = \sum_{i=1}^{\kappa} A_i\bar{r}_i / \sum_{i=1}^{\kappa} A_i$$

where A_i is the area of the ith zone. The variance of $\bar{r}_\omega$ is estimated by

$$S^2_1 = \left(\sum_{i=1}^{\kappa} A_i^2\bar{r}_i/n_i\right)/\left(\sum_{i=1}^{\kappa} A_i\right)^2 \qquad (7.5)$$

The optimal allocation of the n stations to the κ zones is obtained by choosing n_1, $n_2 \ldots n_k$, which minimizes var $(\bar{r}_\omega)$ subject to the condition that $n_1 + n_2 + \ldots n_k = n$. It is easy to show that the values of n_i are given by

$$n_i = n(A_i\bar{r}_i)/\sum_{i=1}^{\kappa} A_i\bar{r}_i \quad (i = 1,2\ldots k) \qquad (7.6)$$

This formula shows that the number of stations allocated to the ith zone is proportional to the area of the zone multiplied by the mean bacterial density of that zone. It should be noted that when r_i are approximately equal, the allocation is approximately proportional to the area of the zone. Substituting the value of n_i from 7.6 into 7.5 gives the estimate of the minimum value of the variance of $\bar{r}_w$ as

$$S^2{}_2 = (\sum_{i=1}^{\kappa} A_i\bar{r}_i)/n\sum_{i=1}^{\kappa} A_i \qquad (7.7)$$

To determine the confidence interval for the bacterial density when the optimal allocation is used, it is usually assumed that $\bar{r}_\omega$ is normal, which is justified using the Central limit theorem with variance given by 7.7. Hence the number of stations n that need to be sampled to detect a pre-specified difference, $\bar{d} = (r_\omega - \mu)$, between the estimated density r_ω and the true but unknown density μ of the lake can be determined. For example, the value of n is

$$n \leq (1.96)^2(\sum_{i=1}^{\kappa} A_i\bar{r}_i/\sum_{i=1}^{\kappa} A_i)/d^2 \qquad 7.8$$

for detecting the difference d using the 5% significance level.

APPLICATIONS

The data discussed in El-Shaarawi *et al.* (1981) are used to illustrate the procedures given by El-Shaarawi and Esterby (1981). These data consist of coliform counts which were obtained by the membrane filtration technique, using M. Endo broth (Difco) with 35°C incubation for 24 h. The bacteriological analysis was performed on water samples collected in 1968 at a number of stations in Lake Erie during the following four cruises: (i) May 17 to 24, (ii) July 29 to August 3, (iii) August 31 to September 3, and (iv) November 4 to 10. It was shown in El-Shaarawi *et al.* (1981) that the bacterial counts in the lake had a very high degree of spatial heterogeneity and hence the lake was divided into a number of homogeneous zones (see Figure 4 in El-Shaarawi *et al.*, 1981). In this section, the results of the previous study (El-Shaarawi *et al.*, 1981) are used for (i) optimally allocating the stations to the zones, assuming the same number of stations as that used in the 1968 survey; and (ii) determining the number of stations that must be sampled and their distribution over the lake to obtain exactly the same efficiency as that obtained using the 1968 design. The difference between this and the corresponding number of stations used in 1968 is an indication of the amount of resources saved by adopting the optimal allocation method, which, of course, could have been accomplished if prior information about the bacterial variability of the lake had been available before the 1968 surveys were conducted.

Table 66 gives the arithmetic mean, $\bar{r}_a$, and the areal weighted mean, $\bar{r}_\omega$, of the bacterial density in columns 2 and 3, respectively. The variance actually achieved by this design for r_ω is obtained using Equation 7.5 and is given in column 4. The corresponding estimate for the variance, when the optimal allocation is used, is calculated from Equation 7.7 and is given in column 5. The ratio of the optimal variance to the actual variance gives the efficiency of the 1968 design, which is shown in column 6. This shows that the minimum efficiency is 0.387, which occurs during the May 17 to 24 cruise, while the maximum efficiency of 0.720 occurs during the July 20 to August 3 cruise. Another way of discussing the efficiency is to determine how many stations must be sampled when the optimal allocation is used to obtain the same variance as that obtained using the 1968 design. This can be done by setting the variance S^2_2 of $\bar{r}_\omega$ in Equation 7.7 equal to the actual observed variance and solving for the value of n for each zone. Column 8 of Table 66 gives these values, denoted by $\tilde{n}$, which can be compared with the actual number of stations n used in 1968 as given in column 7.

When the number of stations is specified, formula 7.6 is used to determine the number of stations that should be allocated to each zone. Assume first that the number of stations to be sampled is equal to that used in 1968, and hence this allows the comparison between the optimal allocation and the actual allocation. Secondly, assume that the number of stations to be used gives the same efficiency as that of 1968. The results of these calculations using the zones given in the study by El-Shaarawi et al. (1981) are given in Table 67. Columns headed by n give the 1968 allocations, while columns headed by $\hat{n}$ give the optimal allocation, and the columns headed by $\tilde{n}$ present the allocation which is of the same efficiency as the 1968 design. Generally, Table 67 shows that the optimal design reduces the number of stations in the first zone and increases the number of stations in the zones with high bacterial densities. This can be seen by comparing the values of n with those of $\hat{n}$.

Table 68 and Figure 83 give the number of stations that must be sampled to detect a pre-specified relative difference $d/\bar{r}_\omega$ at the 5% significance level. These show that the number of stations declines very rapidly with increasing $d/\bar{r}_\omega$ and also that the May 17-24 cruise displayed a pattern similar to that of the July 20 to August 3 cruise, whereas the other two cruises are similar. From the figure or the table, it can be seen that 25 stations would be enough to detect a relative difference of 0.15 for any of the four time periods considered, but this number would be smaller for some of the dates.

The application of the methods to the 1968 data of Lake Erie showed that the design used was not efficient. This is due to the high degree of the spatial variability in the lake, and hence more stations should be allotted to areas with large variabilities. Furthermore, if the efficiency of the future design has to be the same as that of 1968, a great reduction in the number of stations to be sampled is possible. Also, the question of specifying the number of stations needed to detect a specific relative difference is addressed, and it is found that it is possible to use only 25 stations and still be capable of detecting a relative difference equal to 0.15.

GENERALIZATION

This section gives a generalization of the problem of sampling design when the lake is divided into κ homogeneous zones. There are two aspects involved in this generalization. The first is that the distribution of the variable of interest is unknown and the second is that there is a cost associated with the data collection. Suppose that using historical data it is possible to obtain an estimate of the standard deviation in each zone and let these be $s_1, s_2 \ldots s_\kappa$. Furthermore, suppose that the object is to estimate the areal weighted mean of the limnological variable. Then any sample design which allocates n_1 stations to the first zone, n_2 stations to the second zone... and n_κ stations to the κth zone will result in a particular estimate for the areal weighted mean $\bar{x}_\omega$. This estimate is given by

$$\bar{x}_\omega = (\sum_{i=1}^{\kappa} A_i \bar{x}_i)/\sum_{i=1}^{\kappa} A_i \qquad (7.9)$$

where $\bar{x}_i$ is the mean of n_i observations from the ith zone ($i=1,2...\kappa$). Assume further that visiting a sampling station in the ith zone will result in a cost $c_i(i=1,2...\kappa)$. This means that it is assumed that the cost of obtaining an observation is the same for sampling stations within a particular zone, but the cost differs from zone to zone. Hence under this assumption the total cost of conducting the data collection is

$$C = (n_1 c_1 + n_2 c_2 + \ldots + n_\kappa c_\kappa) \qquad (7.10)$$

The estimated variance of $\bar{x}_\omega$ is

$$s^2 = \text{var}\ (\bar{x}_\omega) = \sum_{i=1}^{\kappa} A_i^2\ (s_i^2/n_i)/(\sum_{i=1}^{\kappa} A_i)^2 \qquad (7.11)$$

Assuming that the total cost is constant and the total number of sampling stations is also constant, that is

$$n = n_1 + n_2 + \ldots + n_\kappa \qquad (7.12)$$

the problem is how to estimate n_i ($i = 1,2\ldots\kappa$) which minimizes 7.11 subject to the conditions given by 7.10 and 7.12. The solution can be obtained numerically. However, if the condition 7.12 is disregarded, then the value of n_i is given as

$$n_i = \frac{CA_i \cdot s_i}{(\sum A_i s_i)\ \sqrt{c_i}}, \quad (i = 1,2\ldots\kappa) \tag{7.13}$$

which shows that the number of stations allocated to the ith zone increases with increasing the standard deviation, the area of the zone and with decreasing cost. Substituting n_i in 7.11 yields the minimum value of the variance under condition 7.10. The optimal variance can then be used to obtain an estimate of the number of stations that must be sampled with a pre-specified precision as given in the previous section.

SAMPLING STRATEGY WHEN THE CONCENTRATION IS A CONTINUOUS FUNCTION

In this case it is assumed that the values of the limnological variable y can be represented as a continuous function of two coordinates x_1 and x_2 which specify the position of the sampling station in the lake (x_1 is the longitude and x_2 is the latitude), i.e.,

$$y = f(x_1, x_2) + \varepsilon \tag{7.14}$$

where f is the assumed form of the function (surface) and ε is a random variable with mean 0 and variance σ^2. El-Shaarawi and Esterby (1981) used a quadratic surface to represent some limnological variables, i.e.,

$$f(x_1,x_2) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_{12} x_1 x_2 + \alpha_{11} x_1^2 + \alpha_{22} x_2^2 = \underline{\alpha}\,\underline{a} \tag{7.15}$$

where $\underline{\alpha}$ is the vector of the unknown parameters α_1, α_2, α_{12}, α_{11} and α_{22} and $\underline{a}$ is the vector with the elements 1, x_1, x_2, $x_1\,x_2$, $x^2{}_1$ and $x^2{}_2$. Let $\hat{\underline{\alpha}}$ be the estimated value of $\underline{\alpha}$ and $\hat{\sigma}^2$ be the estimated value of σ^2 from historical data. Under these conditions it is the objective to use model 7.15 and the estimate of the parameters to determine an appropriate strategy for estimating the areal weighted mean of the limnological variable, which is given by

$$\bar{y}_\omega = \iint_Q f(x_1,x_2)dx_1dx_2 / \iint_Q dx_1dx_2 \tag{7.16}$$

where Q represents the surface area of the lake. Substituting expression 7.15 in 7.16 and integrating yields the estimate $\bar{y}_\omega$ as a function of $\underline{\alpha}$.

Any sample design will result in sampling at a number of locations and will lead to the estimate $\underline{\alpha}^*$ of $\underline{\alpha}$. The variance of $\underline{\alpha}^*$ is a function of the values of x_1 and x_2 and hence the variance of $\bar{y}_\omega$ will also be a function of x_1 and x_2. Then for a fixed value of n the position of x_1 and x_2 which minimizes the variance of $\bar{y}_\omega$ has to be determined numerically by computer.

APPLICATIONS

Because the computer program for performing the analysis described in the previous section is still in the developmental stage, a simple empirical approach is used to study the effect of the choice of the sampling design on estimating the areal weighted mean of the lake. This is done by assuming that an existing sampling pattern represents the "true" population and hence the estimated vector $\underline{\hat{\alpha}}$ is taken as the "true" vector of the population parameters. Then a number of sampling designs are generated from this population. For example, the first design is obtained by taking every second consecutive station, so that the sample design in this case represents half the population. The second design is obtained by taking every third consecutive station and so on for other designs. Figure 84 shows the variance of $\bar{y}_{\omega}$ for temperature against the number of sampling stations when the cruises of 1968 and 1977 are taken as the true population. The date of each cruise is shown on the corresponding variance curve. The figure shows that the variance decreases substantially for designs with the number of stations less than 20. The degree of decrease in the variance is very slight after 20 stations. This suggests that between 20 and 25 stations are adequate for estimating the areal weighted mean of temperature. The same conclusion is reached for chlorophyll a where the results of doing the previous analysis are shown in Figure 85. Furthermore, Figure 86 gives the percentage deviation of the areal weighted mean $\bar{y}_{\omega}$ from the "true" areal weighted mean (i.e., mean based on all stations sampled during the cruise) for chlorophyll a. This shows that the magnitude of the percentage deviation does not exceed 7.5% when a design with 25 sampling stations is used to monitor the lake.

Table 66. Mean, Weighted Mean, Variance, Efficiency, and the Number of Stations for Each Cruise

Date	Areal mean, $\bar{r}_a$	Weighted mean, $\bar{r}_\omega$	Variance		Efficiency	No. of stations	
			Actual	Optimal		Actual n	Estimated $\bar{n}$
May 17-24	6.581	6.848	0.1239	0.0480	0.387	62	24.02
July 29-Aug.3	13.771	11.712	0.1888	0.1362	0.720	61	44.00
Aug.31-Sept.3	5.500	6.010	0.2801	0.1382	0.491	28	13.80
Nov.4-10	5.000	3.730	0.0526	0.0305	0.580	55	32.00

Table 67. Optimal Allocation of Sampling Stations to Lake Erie

	May 17 to 24				July 29 to Aug. 3				Aug. 31 to Sept. 3				Nov. 4 to 10			
	Mean	No. of stations			Mean	No. of stations			Mean	No. of stations			Mean	No. of stations		
Zone	$\bar{r}$	n	$\hat{n}$	$\bar{n}$	$\bar{r}$	n	$\hat{n}$	$\bar{n}$	$\bar{r}$	n	$\hat{n}$	$\bar{n}$	$\bar{r}$	n	$\hat{n}$	$\bar{n}$
I	0.566	43	20.46	7.92	0.750	24	6.10	4.4	1.318	22	12.0	6.0	0.306	36	17.6	10.2
II	7.833	13	11.16	4.32	9.684	19	30.50	22.0	13.600	5	10.8	5.4	5.700	10	12.0	7.0
III	35.500	4	22.94	8.88	25.000	10	11.00	7.9	38.000	1	5.2	2.6	17.140	7	19.5	11.4
IV	72.667	2	7.44	2.88	48.500	8	13.40	9.7					44.000	2	5.9	3.4

Table 68. Relative Deviation and the Corresponding Sample Size Which Is Capable of Detecting Differences at 5% Significance Level

May 17 to 24		July 20-Aug. 3		Aug. 31-Sept. 3		Nov. 4-10	
$d/\bar{r}$	n	$d/\bar{r}$	n	$d/\bar{r}$	n	$d/\bar{r}$	n
0.001	114 325	0.001	319 102	0.002	148 635	0.002	64 424
0.003	18 581	0.002	79 776	0.003	37 158	0.005	16 106
0.004	12 703	0.004	12 764	0.005	16 515	0.008	7 158
0.006	7 145	0.009	3 191	0.007	9 289	0.011	4 026
0.007	4 573	0.013	1 418	0.008	5 945	0.013	2 577
0.015	1 143	0.017	797	0.017	1 486	0.027	644
0.022	508	0.026	354	0.025	660	0.040	286
0.029	285	0.034	199	0.033	371	0.054	161
0.044	127	0.043	127	0.050	165	0.080	71
0.058	71	0.051	88	0.067	93	0.107	40
0.073	45	0.060	65	0.083	59	0.134	25
0.088	31	0.068	50	0.100	41		
0.102	23	0.077	39	0.116	30		
0.117	18	0.085	32	0.133	23		

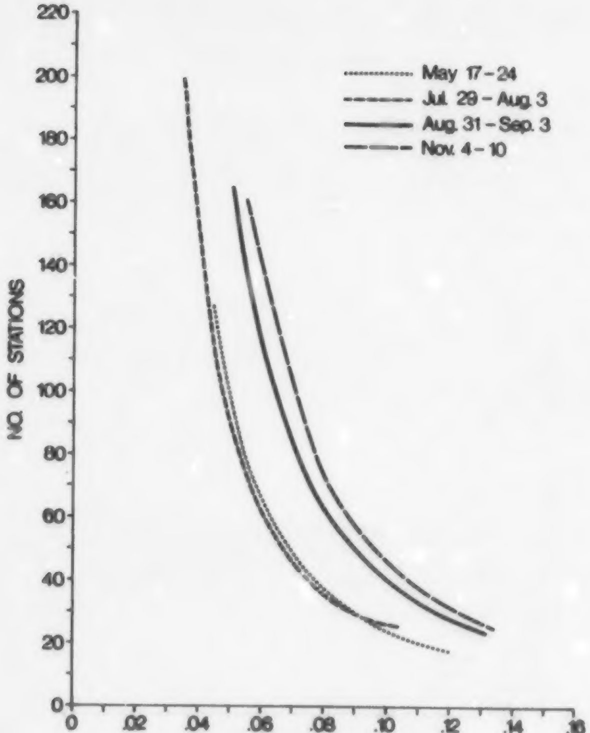


Figure 83. The number of stations needed to detect a difference in the mean at the 5% significance level against relative deviations.

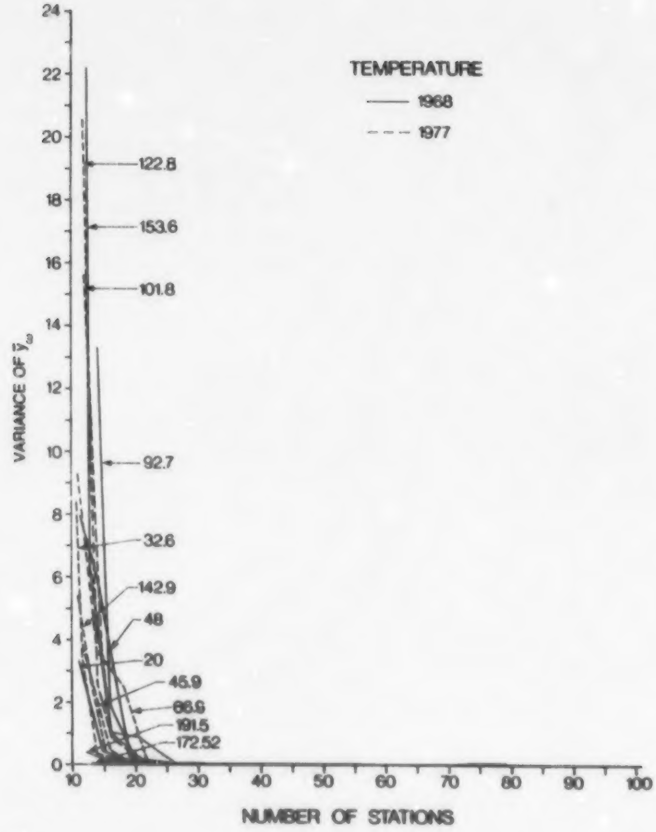


Figure 84. The variance of the mean surface temperature as a function of the number of stations.

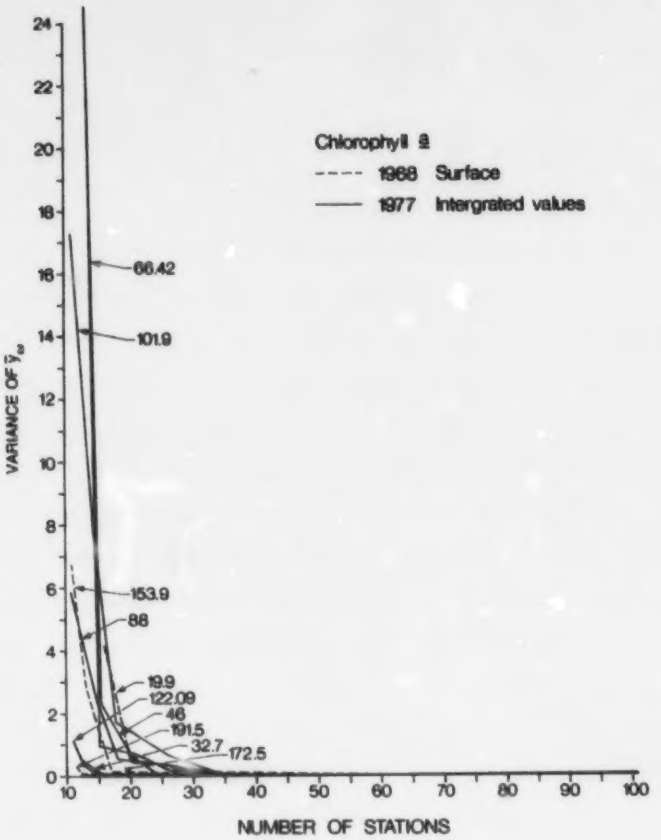


Figure 85. The variance of chlorophyll a as a function of the number of stations.

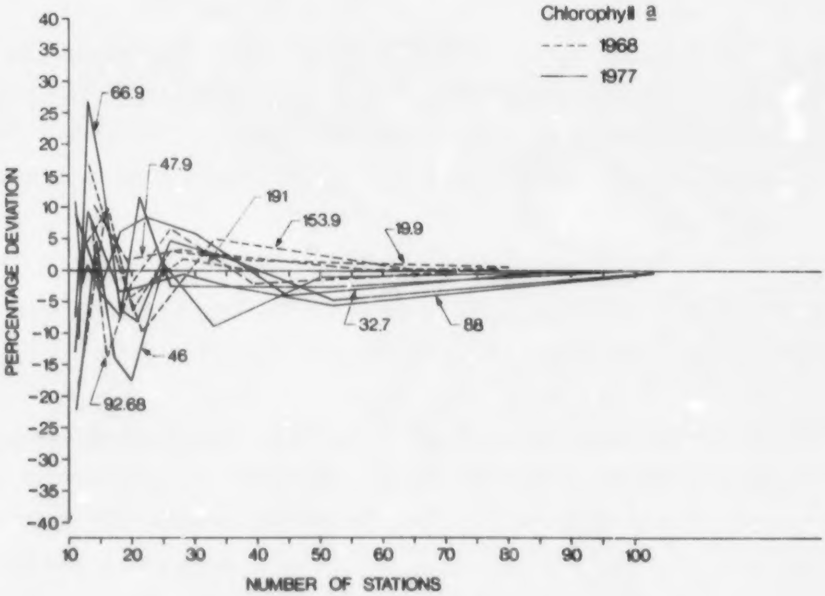


Figure 86. The percentage deviation of the mean chlorophyll a for different sample size.

References

Aitchison, J. and J.A.C. Brown. 1957. The Lognormal Distribution. Cambridge University Press, Cambridge, 176 pp.

Anderson, J.E. 1982. Analysis of dissolved oxygen in Lake Erie (1967-1980) from measured data using cluster analysis. Masters Thesis, Department of Civil Engineering, University of Waterloo, Waterloo, Ontario.

Anderson, J.E., A.H. El-Shaarawi, S.R. Esterby and T.E. Unny. 1984. Spatial and temporal variability of dissolved oxygen in Lake Erie. Chapter 4 of this report.

Baldwin, N.S. and R.W. Saalfield. 1962. Great Lakes Fish. Comm. Tech. Rep. 3, plus supplement, 166 pp.

Ball, G.H. and D.J. Hall. 1967. A clustering technique for summarizing multivariate data. Behav. Sci. 12: 153-157.

Barica, J. 1982. Lake Erie depletion controversy. J. Great Lakes Res. 8(4):719-722

Beeton, A.M. 1963. Limnological survey of Lake Erie, 1959 and 1960. Great Lakes Fish. Comm. Tech. Rep. 6, 35 pp.

Beeton, A.M. 1965. Eutrophication of the St. Lawrence Great Lakes. Limnol. Oceanogr. 10:249-254.

Bierman, V.J., Jr. 1980. A comparison of models developed for phosphorus management in the Great Lakes, pp. 235-255. In Phosphorus Management Strategies for Lakes, R.C. Loehr, C.S. Martin and W. Rast (eds.), Ann Arbor Sci., Ann Arbor, Mich., 490 pp.

Birge, E.A. 1898. Plankton studies on Lake Mendota II. The Crustacea of the plankton from July, 1894, to December, 1896. Trans. Wis. Acad. Sci. Arts Lett. 11:274-448.

Birge, E.A. 1904. The thermocline and its biological significance. Trans. Am. Microsc. Soc. 25:5-33.

Birge, E.A. 1906. Gases dissolved in the waters of Wisconsin lakes. Trans. Am. Fish. Soc. 1906:143-163.

Birge, E.A. 1910. An unregarded factor in lake temperature. Trans. Wis. Acad. Sci. Arts Lett. 16(2):989-1016.

Birge, E.A. 1916. The work of the wind in warming a lake. Trans. Wis. Acad. Sci. Arts Lett. 18:341-391.

Birge, E.A. and C. Juday. 1919. Further limnological observations on the Finger Lakes of New York. Bull. U.S. Bur. Fish. 37.

Birge, E.A. and C. Juday. 1921. The inland waters of Wisconsin. The dissolved gases of the water and their biological significance. Wis. Geol. Nat. Hist. Surv. Bull. XXII.

Brönsted, I.N. and C. Wesenberg-Lund. 1912. Chemisch-physikalische Untersuchungen der dänischen Gewässer nebst Bemerkungen über ihre Bedeutung für unsere Auffassung der Temporalvariationen. Int. Rev. Hydrobiol. 4:251-290.

Brown, R.L., J. Durbin and J.M. Evans. 1975. Techniques for testing the constancy of regression relationships over time. J. R. Statist. Soc., Ser. B, 37, 149-192.

Brownlee, K.A. 1965. Statistical Theory and Methodology in Science and Engineering. John Wiley and Sons, Inc., New York, 590 pp.

Burkholder, J.J. 1960. A survey of the microplankton of Lake Erie. U.S. Fish Wildl. Serv. Spec. Sci. Fish. Rep. No. 334, Washington, D.C., pp. 123-144.

Burns, N.M. 1976a. Oxygen depletion in the Central and Eastern basins of Lake Erie, 1970. J. Fish. Res. Board Can. 33:512-519.

Burns, N.M. 1976b. Temperature, oxygen, and nutrient distribution patterns in Lake Erie, 1970. J. Fish. Res. Board Can. 33:485-511.

Burns, N.M. and F. Rosa. 1981. Oxygen depletion rates in the hypolimnion of central and eastern Lake Erie. A new approach indicates change. Unpublished working draft presented at a workshop on Lake Erie Oxygen Depletion Controversy, December, 2-3, 1981, Canada Centre for Inland Waters, Burlington, Ontario.

Burns, N.M. and C. Ross. 1972a. Oxygen-nutrient relationships within the Central Basin of Lake Erie, pp. 85-119. In Project Hypo, N.M. Burns and C. Ross (eds.), Canada Centre for Inland Waters, p. 6, U.S. EPA Tech. Rep. TS-05-71-208-24, 182 pp.

Burns, N.M. and C. Ross. 1972b. Project Hypo - Discussion of Findings, pp. 120-126. In Project Hypo, N.M. Burns and C. Ross (eds.), Canada Centre for Inland Waters, p. 6, U.S. EPA Tech. Rep. TS-05-71-208-24, 182 pp.

Canada Centre for Inland Waters. 1966-1976. Lake Erie. Limnological Data Reports. Canada Centre for Inland Waters, Burlington, Ontario.

Carr, J.F. 1962. Dissolved oxygen in Lake Erie, past and present. Great Lakes Research Division, University of Michigan, Pub. No. 9:1-14.

Carr, J.F. and J.K. Hiltunen. 1965. Changes in the bottom fauna of western Lake Erie from 1930 to 1961. Limnol. Oceanogr. 10:551-569.

Chapra, S.C. 1980. Application of the phosphorus loading concept to the Great Lakes, pp. 135-152. In Phosphorus Management Strategies for Lakes, R.C. Loehr, C.S. Martin and W. Rast (eds.), Ann Arbor Sci., Ann Arbor, Mich., 490 pp.

Charlton, M.N. 1979. Hypolimnetic oxygen depletion in central Lake Erie: Has there been any change? Sci. Ser. No. 110, Inland Waters Directorate, Environment Canada, Burlington, Ontario, 25 pp.

Charlton, M.N. 1980a. Oxygen depletion in Lake Erie: Has there been any change? Can. J. Fish. Aquat. Sci. 37:72-81.

Charlton, M.N. 1980b. Hypolimnion oxygen consumption in lakes: Discussion of productivity and morphometry effects. Can. J. Fish. Aquat. Sci. 37:1531-1539.

Dambach, C.A. 1969. Changes in the biology of the lower Great Lakes. Bull. Buffalo Soc. Nat. Sci. 25:1-10.

Davies, C.C. 1964. Evidence for the eutrophication of Lake Erie from phytoplankton records. Limnol. Oceanogr. 9:275-283.

Davies, C.C. 1969. Plants in Lake Erie and Ontario, and changes of their numbers and kinds. Bull. Buffalo Soc. Nat. Sci. 25:18-41.

DiToro, D.M. 1980. The effect of phosphorus loadings on dissolved oxygen in Lake Erie, pp. 191-206. In Phosphorus Management Strategies for Lakes, R.C. Loehr, C.S. Martin and W. Rast (eds.), Ann Arbor Sci., Ann Arbor, Mich., 490 pp.

Dobson, H.H. and M. Gilbertson. 1971. Oxygen depletion in the hypolimnion of the Central Basin of Lake Erie, 1929 to 1970. Proc. Great Lakes Res. 14:743-748.

Duda, R. and P. Hart. 1973. Pattern Classification and Scene Analysis. John Wiley and Sons, Inc., New York.

Durbin, J. 1969. Tests for serial correlation in regression analysis based on the periodogram of least squares residuals. Biometrika, 56: 1-15.

El-Shaarawi, A.H. 1982. Sampling strategy for estimating bacterial density in large lakes. Journal francais d'hydrologie, 13:171-187.

El-Shaarawi, A.H. 1984a. Temporal changes in Lake Erie. Chapter 3 of this report.

El-Shaarawi, A.H. 1984b. Dissolved oxygen concentrations in Lake Erie. 2. A statistical model for dissolved oxygen in the Central Basin of Lake Erie. Chapter 5 of this report.

El-Shaarawi, A.H. and S.R. Esterby. 1981. Analyse de régression de la variation spatiale d'une caractéristique limnologique. Eau du Québec, 14(3).

El-Shaarawi, A.H. and R.E. Kwiatkowski. 1977. A model to describe the inherent spatial and temporal variability of parameters in Lake Ontario, 1974. J. Great Lakes Res. (3-4), pp. 177-183.

El-Shaarawi, A.H. and W.O. Pipes. 1982. Enumeration and statistical inferences, Chapter 3. In Bacterial Indicators of Pollution, W.O. Pipes (ed.), CRC Press, Boca Raton, Florida, 174 pp.

El-Shaarawi, A.H. and K.R. Shah. 1978. Statistical procedures for classification of a lake. Sci. Ser. No. 86, Inland Waters Directorate, Environment Canada, Burlington, Ontario.

El-Shaarawi, A.H., S.R. Esterby and B.J. Dutka. 1981. Bacterial density water determined by Poisson or negative binomial distributions. Appl. Environ. Microbiol., pp. 107-116.

Environment Canada. 1975. Analytical Methods Manual. Water Quality Branch, Inland Waters Directorate, Ottawa.

Environment Canada. 1979. Water Quality Sourcebook, A Guide to Water Quality Parameters, Addendum. Inland Waters Directorate, Water Quality Branch, Ottawa.

Esterby, S.R. 1982. Fitting probability distributions to bacteriological data: considerations for regulations and guidelines. Journal francais d'hydrologie, 13:189-203.

Esterby, S.R. and A.H. El-Shaarawi. 1981. Characterization of spatial and temporal variability of water quality parameters. Summaries of Conference Presentations, Environmentrics 81, Alexandria, Virginia, pp. 156-157.

Fay, L.A. and C.E. Herdendorf. 1981. Lake Erie water quality: Assessment of 1980 open lake conditions and trends for the preceding decade. CLEAR Technical Report No. 219, Ohio State University, Center for Lake Erie Area Research, Columbus, Ohio.

Feller, W. 1957. An Introduction to Probability Theory and its Applications. John Wiley and Sons, Inc., New York, 509 pp.

Fish, C.J. 1960. Limnological survey of eastern and central Lake Erie, 1928 to 1929. U.S. Fish Wildl. Serv. Spec. Sci. Fish Rep. No. 334, Washington, D.C., p. 198.

Fisher, R.A. 1941. The negative binomial distribution. Ann. Eugen. 11:182-187.

Fisher, R.A., H.G. Thornton and W.A. MacKenzie. 1922. The accuracy of the plating method of estimating the density of bacterial populations. Ann. Appl. Biol. 9:325-359.

Hamblin, P.F. 1971. Circulation and water movement in Lake Erie. Sci. Ser. No. 7, Dep. Energy, Mines and Resources, Ottawa, 50 pp.

Harris, G. and R.A. Vollenweider. 1982. Paleolimnological evidence of early eutrophication in Lake Erie. Can. J. Fish. Aquat. Sci. 39:618-626.

Health and Welfare Canada. 1979. Guidelines for Canadian Drinking Water Quality. Canadian Government Publishing Centre, Supply and Services Canada, Hull, Quebec.

Hoppe-Seyler, F. 1895. Über die Verteilung absorbierter Gase im Wasser des Bodensees und ihre Beziehungen zu den in ihm lebenden Tierem und Pflanzen. Schr. Ver. Gesh. Bodensees. 24:39-48.

Hutchinson, G.E. 1957. A Treatise on Limnology. Vol. I, John Wiley and Sons, Inc., New York, 1115 pp.

Hutchinson, G.E. 1967. A Treatise on Limnology. Vol. II. Introduction to lake biology and the limnoplankton. John Wiley and Sons, Inc., New York, 1115 pp.

International Joint Commission. 1975. Great Lakes Water Quality, 1974. 3rd Annual Report of the Great Lakes Water Quality Board to the International Joint Commission, Windsor, Ontario, 170 pp.

International Joint Commission. 1976a. Great Lakes Water Quality, 1975. 4th Annual Report of the Great Lakes Water Quality Board to the International Joint Commission, Windsor, Ontario, 72 pp.

International Joint Commission. 1976b. Great Lakes Water Quality, 1975. Appendix B, Surveillance Subcommittee Report, 255 pp.

International Joint Commission. 1977. Great Lakes Water Quality, 1976. Appendix B, Surveillance Subcommittee Report, 134 pp.

International Joint Commission. 1978a. Great Lakes Water Quality, 1977. 6th Annual Report of the Great Lakes Water Quality Board to the International Joint Commission, Windsor, Ontario, 89 pp.

International Joint Commission. 1978b. Great Lakes Water Quality, 1977. Appendix B, Surveillance Subcommittee Report, 110 pp.

International Joint Commission. 1979a. Great Lakes Water Quality, 1978. 7th Annual Report of the Great Lakes Water Quality Board to the International Joint Commission, Windsor, Ontario, 110 pp.

International Joint Commission. 1979b. Great Lakes Water Quality, 1978. Appendix B. Surveillance Subcommittee Report, 117 pp.

International Joint Commission. 1980a. 1980 Report on Great Lakes Water Quality. Appendix, 82 pp.

International Joint Commission. 1980b. Great Lakes Water Quality, 1979. 8th Annual Report of the Great Lakes Water Quality Board to the International Joint Commission, Windsor, Ontario, 150 pp.

International Lake Erie Water Pollution Board and the International Lake Ontario - St. Lawrence Water Pollution Board. 1969. Report to the International Joint Commission, Vol. 1, Summary, 150 pp.

Kalbfleisch, J.D. and D.A. Sprott. 1973. Marginal and conditional likelihoods. Sankhya, Ser. A, 35:311-328.

Lam, D.C.L., W.M. Schertzer and A.S. Fraser. 1983. Simulation of Lake Erie water quality responses to loading and weather variations. Sci. Ser. No. 134, Inland Waters Directorate, Environment Canada, Burlington, Ontario.

Leach, J.H. and S.J. Nepsy. 1976. The fish community in Lake Erie. J. Fish. Res. Board Can. 33:622-638.

Lee, G.F., R.A. Jones and W. Rast. 1980. Availability of phosphorus to phytoplankton and its implications for phosphorus management strategies. pp. 259-308. In Phosphorus Management Strategies for Lakes, R.C. Loehr, C.S. Martin and W. Rast (eds.), Ann Arbor Sci., Ann Arbor, Mich., 490 pp.

Lund, J.W.G. 1965. The ecology of the fresh water phytoplankton. Biol. Rev. 40:231-293.

Nicholls, K.H. and P.J. Dillon. 1978. An evaluation of phosphorus - chlorophyll-phytoplankton relationships for lakes. Int. Rev. Gesamten Hydrobiol. 63:141-159.

Northouse, R.A. and F.R. Fromm. 1976. Class: A nonparametric clustering algorithm. Pattern Recognition, 8:107-114.

Pipes, W.O., P. Ward and S.H. Ahn. 1977. Frequency distributions for coliform bacteria in water. J. Am. Water Works Assoc. 69:664-668.

Phosphorus Management Strategies Task Force. 1980. Phosphorus Management for the Great Lakes. Final Report, 129 pp.

Rao, S.S. and B.K. Burnison. 1976. Bacterial distributions in Lake Erie (1967, 1970). J. Fish. Res. Board Can. 33: 547-580.

Reitz, R.D. 1973. Distribution of Phytoplankton and Coliform Bacteria in Lake Erie. Ohio Environmental Protection Agency, Division of Surveillance.

Richards, R.P. 1981. Historical trends in water chemistry in the U.S. nearshore, central basin, Lake Erie. Water Quality Laboratory Tech. Rep. Ser. No. 15, Heidelberg College, Tiffin, Ohio, 31 pp.

Rockwell, D.C., C.J. Marion, M.F. Palmer, D.S. DeVault and R.J. Bowden. 1980. Environmental trends in Lake Michigan. Proceedings of the 1979 Conference on Phosphorus Management Strategies for Lakes, R.C. Loehr, C.S. Martin and W. Rast (eds)., Ann Arbor Sci., Ann Arbor, Mich., pp. 91-134.

Rosa, F. and N.M. Burns. 1981. Oxygen depletion rates in the hypolimnion of central and eastern Lake Erie. A new approach indicates change. Presented at Workshop on Central Basin Oxygen Depletion, Dec. 2-3, 1981, Canada Centre for Inland Waters, Burlington, Ontario.

Ruttner, F. 1952. Fundamentals of Limnology. Translated by D.G. Frey and F.E.J. Fry, University of Toronto Press, Toronto, 295 pp.

Simons, T.J. 1976. Continuous dynamical computations of water transports in Lake Erie for 1970. J. Fish. Res. Board Can. 33:371-384.

Slater, R.W. and G.E. Bangay. 1980. Action taken to control phosphorus in the Great Lakes. Proceedings of the 1979 Conference on Phosphorus Management Strategies for Lakes, R.C. Loehr, C.S. Martin and W. Rast (eds.), Ann Arbor Sci., Ann Arbor, Mich., pp. 13-26.

Sly, P.G., 1976. Lake Erie and its basin. J. Fish. Res. Board Can. 33:355-370.

Smirnov, V.I. 1964. Linear Algebra. Pergamon Press, London, 324 pp.

Snedecor, G.W. and W.G. Cochran. 1967. Statistical Methods. The Iowa State University Press, Ames, Iowa, 593 pp.

Sprott, D.A. and J.G. Kalbfleisch. 1965. Use of the likelihood function in inference. Psychol. Bull. 64:15-22.

Thienemann, A. 1915. Physikalische und chemische untersuchungen in den Meeren der Eifel. Verh. naturh. Ver. preuss. Rheinl. 71:281-389.

Thienemann, A. 1928. Der Sauerstoff in eutrophen und oligotrophen Seen. Die Binnengewässer, Vol. 4. E. Schweizerbart'sche Verlagsbuchhandlung, Stuttgart, Germany, 175 pp.

Thomann, R.V. and J.S. Segna. 1980. Dynamic phytoplankton-phosphorus model of Lake Ontario: ten year verification and simulations, pp. 153-170. In Phosphorus Management Strategies for Lakes, R.C. Loehr, C.S. Martin and W. Rast (eds.), Ann Arbor Sci., Ann Arbor, Mich., 490 pp.

Thomas, N.A., A. Robertson and W.C. Songzogni. 1980. Review of control objectives, new target loads and input controls, pp. 61-90. In Phosphorus Management Strategies for Lakes, R.C. Loehr, C.S. Martin and W. Rast (eds.), Ann Arbor Sci., Ann Arbor, Mich., 490 pp.

Vollenweider, R.A. and L.L. Janus. 1981. Statistical models for predicting hypolimnetic oxygen depletion rates. Presented at Workshop on Central Basin Oxygen Depletion, Dec. 2-3, 1981, Canada Centre for Inland Waters, Burlington, Ontario.

Vollenweider, R.A., W. Rast and J. Kerekes. 1980. The phosphorus loading concept and Great Lakes eutrophication, pp. 207-234. In Phosphorus Management Strategies for Lakes, R.C. Loehr, C.S. Martin and W. Rast (eds.), Ann Arbor Sci., Ann Arbor, Mich., 490 pp.

Wetzel, R.G. and G.E. Likens. 1979. Limnological Analysis. W.B. Saunders Co., Toronto.

Zar, H. 1980. Point source loads of phosphorus to the Great Lakes, pp. 27-36. In Phosphorus Management Strategies for Lakes, R.C. Loehr, C.S. Martin and W. Rast (eds.), Ann Arbor Sci., Ann Arbor, Mich., 490 pp.

Appendix

Examination of Linear Regression Models by Residual Analysis

Examination of Linear Regression Models by Residual Analysis

INTRODUCTION

In this appendix the adequacy of linear regression models is examined. The approach used is based on the analyses of residuals which are the differences between the observed values and their corresponding estimated values. Several functions of residuals are presented and their uses in detecting the inadequacy of regression models are explained.

LINEAR REGRESSION MODELS

Let $\underline{y}$ be the column vector $(y_1, y_2 \ldots y_n)$ of n random variables which is generated from the linear model

$$\underline{y} = A\underline{\Theta} + \underline{\varepsilon} \tag{A-1}$$

where A is a matrix of order (n x k) and of full rank k, $\underline{\Theta}$ is a vector of k unknown parameters and $\underline{\varepsilon}$ is a vector of n unobservable independent normally distributed random variables with mean $\underline{0}$ and variance σ^2. It is well known that the best unbiased linear estimate (BLUE) of $\underline{\Theta}$ is

$$\underline{\hat{\Theta}} = (A'A)^{-1} A'\underline{y} \tag{A-2}$$

where for any matrix B, the matrices B' and B^{-1} represent the transpose of B and the inverse of B (when it exists), respectively. The estimated value of $\underline{y}$ is

$$\hat{\underline{y}} = A\hat{\theta} \tag{A-3}$$

using the relation A-2, Equation A-3 can be rewritten as

$$\hat{\underline{y}} = P\underline{y} \tag{A-4}$$

where $p = A(A'A)^{-1}A'$ is the projection matrix. The matrix P is idempotant (i.e., $P^2 = P$) and it has rank k. The vector of residuals is

$$\begin{aligned} \underline{R} &= \underline{y} - \hat{\underline{y}} \\ &= (I - P)\underline{y} \end{aligned} \tag{A-5}$$

where I is the unit matrix. The residual sum of squares is

$$\begin{aligned} S &= \underline{R}'\underline{R} \\ &= \underline{y}'(I - P)\underline{y} \end{aligned} \tag{A-6}$$

The variance σ^2 is estimated by

$$\hat{\sigma}^2 = S/(n-k) \tag{A-7}$$

SOME PROPERTIES OF THE RESIDUALS

The sufficiency of $\underline{\hat{\theta}}$ and $\hat{\sigma}^2$ for the estimation of the parameters $\underline{\theta}$ and σ^2 indicates that the vector of residuals contains all the information for the examination of the adequacy of model A-1 especially the distributional assumption of $\underline{\varepsilon}$. Hence, it is appropriate to investigate the relationship between $\underline{\varepsilon}$ and $\underline{R}$. Since I - P and k are orthogonal, we have

$$\underline{R} = (I - P)\underline{\varepsilon} \tag{A-8}$$

and

$$\underline{R}'\underline{R} = \underline{\varepsilon}'(I - P)\underline{\varepsilon} \tag{A-9}$$

From A-8 the elements of $\underline{R}$ can be expressed as n linear combinations of the elements of $\underline{\varepsilon}$. However, since I - P is not of full rank but of rank (n - k), these linear combinations are not unique. Under the assumptions of $\underline{\varepsilon}$ we have

$$E(\underline{R}) = \underline{0} \tag{A-10}$$

$$\text{var}\ (\underline{R}) = (I - P)\ \sigma^2 \tag{A-11}$$

and the distribution of $\underline{R}$ is multivariate normal with mean and variance covariance matrix given by A-10 and A-11, respectively. This shows that the elements of $\underline{R}$ are correlated and this correlation is solely a function of the design matrix A. Under certain conditions on the elements of A and for large values of n and small values of k, the correlation between the elements of $\underline{R}$ is very small and can be disregarded, i.e. the elements of P are set to zero in A-11. Hence

$\underline{R}/\sigma$ is taken as approximately normal with mean $\underline{0}$ and variance-covariance matrix I.

As an example, consider the model

$$\underline{y} = \underline{1}\theta + \underline{\varepsilon}$$

where $\underline{1}$ is a unit vector and θ is the mean of $\underline{y}$. In this case $k = 1$ and $P = 1\ 1'/n$. Hence the off-diagonal elements of P are all equal to $1/n$, so that for large n the correlation between the elements of $\underline{R}$ can be disregarded.

Another approach that is useful especially for moderate and small values of n is to transform $\underline{R}$ into a set of (n-k) orthogonal residuals. A formula by Jacobi (Smirnov, 1964, p. 130) gives a convenient method for doing this.

JACOBI'S FORMULA

Let $Q = \underline{x}'C\underline{x}$ be a quadratic form and set $\underline{Z} = C\underline{x}$. Define the square matrix C_k as the matrix which consists of the first k rows and k columns of C, M_k is the matrix which is formed from C_k replacing the last column of C_k by $\underline{Z}_k$ where $\underline{Z}_k$ is the vector of length k and its elements are the first k components of Z. Let

$$\Delta_0 = 0,\ \Delta_1 = a_{11},\ \Delta_k = \left|C_k\right| \qquad (k = 2\ldots n)$$

$$X_1 = Z_1,\ X_k = \left|M_k\right| \qquad (k = 2,3\ldots n)$$

where Δ_k and X_k are the determinants of C_k and X_k, respectively. If the rank of the matrix C equals r and the determinants Δ_1, $\Delta_2 \ldots \Delta_r$ are non-vanishing, then Jacobi's formula is

$$Q = \underline{x}'C\underline{x} = \sum_{k=1}^{r} \frac{X^2_k}{\Delta_k \Delta_{k-1}}$$

where the linear forms X_k $(k = 1,2 \ldots r)$ are linearly independent.

Applications of Jacobi's formula give the following set

$$w_i = \{y_i - \underline{a}'_i \hat{\underline{\theta}}_{i-1}\} / \sqrt{1 + \underline{a}'_i (A'_{i-1} A_{i-1})^{-1} \underline{a}_i} \quad (i = k+1, k+2 \ldots n)$$

of orthogonalized residuals, where $\hat{\underline{\theta}}_i$ is the least squares estimate of $\underline{\theta}$ using only the first i observations, $\underline{a}'_i$ is the ith row of A, and A_i is the matrix of the first i rows of A.

The normality of ε_i can then be tested by testing the normality of w_i. If $(n-k)$ is large, then the observed frequency histogram of w_i can be tested for normality using the conventional x^2 method. If $(n-k)$ is small, graphical methods such as P - P plot or Q - Q plot are adequate for detecting the departure from normality. For testing that $E[\varepsilon_i] = 0$ and $Var[\varepsilon_i] = \sigma^2$, the following two methods are used: (i) the cumulative residual CUSUM and (ii) the cumulative squared residual CUSUMS. In the following the use of these methods is described.

THE CUSUM TEST

The test statistics are defined as

$$C_t = \frac{1}{\sigma} \sum_{j=k+1}^{t} w_j \qquad t = k+1 \ldots n \qquad \text{(A-12)}$$

under the null hypothesis

$$H : \quad E(\varepsilon_i) = 0 \qquad i = 1,2 \ldots n$$

the set of random variables C_{k+1}, $C_{k+2} \ldots C_n$ are approximately distributed as multivariate normal with mean $\underline{0}$. The elements of the variance-covariance matrix of C_t ($t = k+1 \ldots n$) are given by

$$\mathrm{var}(C_t) = 1 \qquad t = k+1 \ldots n$$

and

$$\mathrm{cov}(C_t, C_{t+r}) = t-k$$

Brown, Durbin and Evans (1975) showed the critical values $C_\alpha(t)$ of C_t are determined by

$$C_\alpha(t) = h_\alpha \sqrt{n-k} + z \cdot h_\alpha \cdot \frac{t-k}{\sqrt{n-k}} \qquad t=k+1 \ldots n$$

where h_α is the root of the equation

$$1 - \Phi(3h_\alpha) + \Phi(h_\alpha)\, e^{-2h^2 \alpha} = \alpha/2$$

where $\Phi(h_\alpha) = \int_{-\infty}^{h_\alpha} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$

THE CUSUMS TEST

A related test procedure is suggested by Brown, Durbin and Evans (1975) for testing the constancy of the variances of $\varepsilon_1, \varepsilon_2 \ldots \varepsilon_n$ which is given by

$$CS_t = sw_t / sw_n \qquad t = k+1 \ldots n$$

where

$$sw_r = \sum_{j=k+1}^{r} w_j^2$$

under the null hypothesis we have

$$E(CS_t) = (t-k)/(n-k) \cdot t = k+1 \ldots n$$

Critical values $CS_\alpha(t)$ are used to assess the significance of departures of the CS_t's from their expected values simultaneously. These critical values are given by Durbin (1969).

POOR

QUALITY

ORIGINAL